Australian Government

Patent Office
Canberra

I, JULIE BILLINGSLEY, TEAM LEADER EXAMINATION SUPPORT AND SALES hereby certify that annexed is a true copy of the Provisional specification in connection with Application No. 2004901931 for a patent by COCHLEAR LIMITED as filed on 08 April 2004.

COMMONWEALTH OF AUSTRALIA
AUSTRALIA
PATENT OFFICE

WITNESS my hand this
Fourth day of June 2004

J. Billingsley

JULIE BILLINGSLEY
TEAM LEADER EXAMINATION
SUPPORT AND SALES

MULTI-MICROPHONE ADAPTIVE NOISE REDUCTION TECHNIQUES FOR SPEECH ENHANCEMENT

I. Background

In speech communication applications, such as teleconferencing, hands-free telephony and hearing aids, the presence of background noise and/or reverberation may significantly reduce the intelligibility of the desired speech signal. This stems from the large distance between the speaker and the microphone(s). Hence, the use of a noise reduction algorithm is necessary. Multi-microphone systems exploit spatial information in addition to temporal and spectral information of the desired signal and noise signal and are thus preferred to single microphone procedures (such as spectral subtraction). Because of aesthetical reasons, multi-microphone techniques for e.g., hearing aid applications go together with the use of small-sized arrays. Considerable noise reduction can be achieved with such arrays, but at the expense of an increased sensitivity to errors in the assumed signal model such as microphone mismatch, reverberation, ... [1, 2] In hearing aids, microphones are rarely matched in gain and phase. In [3], e.g., gain and phase differences between microphone characteristics of up to 6 dB and 10°, respectively, have been reported.

A widely studied multi-channel adaptive noise reduction algorithm is the *Generalized Sidelobe Canceller* (GSC) [2]-[11], depicted in Figure 1. The GSC consists of a fixed, spatial pre-processor, which includes a fixed beamformer and a blocking matrix, and an adaptive stage based on an Adaptive Noise Canceller (ANC) [12]. The ANC minimizes the output noise power while the blocking matrix should avoid speech leakage into the noise references. The standard GSC assumes the desired speaker location, the microphone characteristics and positions to be known, and reflections of the speech signal to be absent. If these assumptions are fulfilled, it provides an undistorted enhanced speech signal with minimum residual noise. However, in reality these assumptions are often violated, resulting in so-called speech leakage and hence speech distortion. To limit speech distortion, the ANC is adapted during periods of noise only [7, 10, 13]. When used in combination with small-sized arrays, e.g., in hearing aid applications, an additional robustness constraint [9, 10, 14, 15] is required to guarantee performance in the presence of small errors in the assumed signal model, such as microphone mismatch [16, 17]. A widely applied method consists of imposing a Quadratic Inequality Constraint to the ANC (QIC-GSC) [10, 14, 15, 18, 19]. For LMS updating, the Scaled Projection Algorithm (SPA) [14] is a simple and effective technique that imposes this constraint. However, the QIC-GSC goes at the expense of less noise reduction [17].

In [20], a *Multi-channel Wiener Filtering* (MWF) technique has been proposed that provides a Minimum Mean Square Error (MMSE) estimate of the desired signal portion in one of the received microphone signals [21]-[24]. In contrast to the ANC of the GSC, the MWF is able to take speech distortion into account in its optimization criterion. The MMSE optimization criterion of the MWF can also be generalized to allow for a trade-off between speech distortion and noise reduction. We will refer to this generalization as Speech Distortion Weighted MWF (SDW-MWF). The MWF technique is uniquely based on estimates of the second order statistics of the recorded speech signal and the noise signal. A robust speech detection is thus (again) needed. In contrast to the GSC, the MWF does not make any a priori assumptions about the signal model so that no or a less severe robustness constraint is needed to guarantee performance when used in combination with small-sized arrays [16, 17]. Especially in complicated noise scenarios such as multiple noise sources

or diffuse noise, the MWF outperforms the GSC, even when the GSC is supplemented with a robustness constraint [17].

In [20, 21], the implementation of the MWF is based on a Generalized Singular Value Decomposition (GSVD) of an input data matrix and a noise data matrix. A cheaper alternative based on a QR Decomposition (QRD) has been proposed in [22]. A subband implementation [23] results in improved intelligibility at a significantly lower cost compared to the fullband approach. However, in contrast to the GSC and the QIC-GSC [14], no efficient, cheap stochastic gradient based implementation of the (SDW-)MWF, which avoids the use of expensive matrix computations, is available yet. In [25], an LMS based algorithm for the MWF has been developed. The algorithm needs recordings of calibration signals. Since room acoustics, microphone characteristics and the location of the desired speaker change over time, frequent re-calibration is required, making this approach cumbersome and expensive. In [26], an LMS based SDW-MWF has been proposed that avoids the need for calibration signals. The algorithm however relies on some independence assumptions that are not necessarily satisfied, resulting in degraded performance w.r.t. matrix-based implementations.

II. Summary

In the *present invention*, we establish a generalized multi-channel noise reduction scheme, referred to as *Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filter (SP-SDW-MWF)*, that encompasses the GSC and the MWF as extreme cases. In addition, the scheme allows for in-between solutions such as the *Speech Distortion Regularized GSC* (SDR-GSC). The generalized scheme, depicted in Figure 3, consists of a fixed, spatial pre-processor and an adaptive stage that is based on an SDW-MWF, hence the name *Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener filter (SP-SDW-MWF)*.

The SP-SDW-MWF adds robustness against signal model errors to the GSC by taking speech distortion explicitly into account in the design criterion of the adaptive stage. The SP-SDW-MWF is an alternative technique to the widely studied QIC-GSC to decrease the sensitivity of the GSC to signal model errors such as microphone mismatch, reverberation, ... A parameter μ is incorporated in the SP-SDW-MWF that allows for a trade-off between speech distortion and noise reduction. Focussing all attention towards speech distortion (i.e., setting $\mu = 0$) results in the output of the fixed beamformer. In noise scenarios with very low Signal-to-Noise Ratio (SNR), e.g., -10 dB, a fixed beamformer may be preferred. Adaptivity can then be easily reduced or excluded in the SP-SDW-MWF by decreasing the parameter μ to 0. Compared to the widely studied QIC-GSC, the SP-SDW-MWF achieves a better noise reduction performance for a given maximum allowable speech distortion level.

In [22, 27] recursive implementations of the (SDW-)MWF have been proposed based on a GSVD or QR decomposition. A subband implementation [28] results in improved intelligibility at a significantly lower cost compared to the fullband approach. These techniques can be extended to implement the SP-SDW-MWF [29]. However, in contrast to the GSC and the QIC-GSC [14], no cheap stochastic gradient based

implementation of the SP-SDW-MWF is available. In the present invention, we propose time-domain and frequency-domain stochastic gradient implementations of the SP-SDW-MWF that preserve the benefit of the matrix-based SP-SDW-MWF over QIC-GSC.

Below, the different embodiments of the present invention are described.

A *first embodiment* proposes a *Speech Distortion Regularized GSC* (SDR-GSC). A new design criterion is developed for the adaptive stage of the GSC: the ANC design criterion is supplemented with a regularization term that limits speech distortion due to signal model errors. In the SDR-GSC, a parameter μ is incorporated that allows for a trade-off between speech distortion and noise reduction. Focussing all attention to noise reduction, results in the standard GSC, while, on the other hand, focussing all attention towards speech distortion results in the output of the fixed beamformer. In noise scenarios with low SNR, adaptivity in the SDR-GSC can be easily reduced or excluded by increasing attention towards speech distortion, i.e., by decreasing the parameter μ to 0. The SDR-GSC is an alternative technique to the QIC-GSC to decrease the sensitivity of the GSC to signal model errors such as microphone mismatch, reverberation, ... In contrast to the QIC-GSC, the SDR-GSC shifts emphasis towards speech distortion when the amount of speech leakage grows. In the absence of signal model errors, the performance of the GSC is preserved. As a result, a better noise reduction performance is obtained for small model errors, while guaranteeing robustness against large model errors.

In a *second embodiment,* we further improve the noise reduction performance of the SDR-GSC by adding an extra adaptive filtering operation w_0 on the speech reference signal. We refer to this generalized scheme as *Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filter* (SP-SDW-MWF). The SP-SDW-MWF is depicted in Figure 3 and encompasses the MWF [20] as a special case. Again, a parameter μ is incorporated in the design criterion to allow for a trade-off between speech distortion and noise reduction. Focussing all attention to speech distortion, results in the output of the fixed beamformer. Also here, adaptivity can be easily reduced or excluded by decreasing μ to 0. It is shown that -in the absence of speech leakage and for infinitely long filter lengths- the SP-SDW-MWF corresponds to a cascade of a SDR-GSC with a SDW single channel Wiener postfilter (SDW-SWF) [30] and thus outperforms the SDR-GSC. In the presence of speech leakage, the SP-SDW-MWF with w_0 tries to preserve its performance: compared to a SDR-GSC (with SDW-SWF postfilter), the SP-SDW-MWF then contains extra filtering operations that compensate for the performance degradation of the SDR-GSC (with SDW-SWF) due to speech leakage (see also Figure 4). In contrast to the SDR-GSC (and thus also the GSC), performance does not degrade due to microphone mismatch. In [22, 27] recursive implementations of the (SDW-)MWF have been proposed based on a GSVD or QR decomposition. A subband implementation [28] results in improved intelligibility at a significantly lower cost compared to the fullband approach. These techniques can be extended to implement the SDR-GSC and, more generally, the SP-SDW-MWF.

In a *third embodiment,* we propose cheap *time-domain and frequency-domain stochastic gradient implementations* of the SDR-GSC and SP-SDW-MWF. Starting from the design criterion of the SDR-GSC, or more generally, the SP-SDW-MWF, we derive a time-domain stochastic gradient algorithm. In addition,

we modify the LMS based algorithm [26] so that it applies to the SP-SDW-MWF. To increase convergence and reduce complexity, a frequency-domain implementation has been proposed. Both, the stochastic gradient and LMS based algorithm suffer from a large excess error when applied in highly time-varying noise scenarios. We show that the excess error in the stochastic gradient algorithm is reduced by applying a low pass filter to the part of the gradient estimate that limits speech distortion. The low pass filtering avoids a highly time-varying distortion of the desired speech component while not degrading the tracking performance needed in time-varying noise scenarios. The stochastic gradient SP-SDW-MWF outperforms the LMS based algorithm, while complexity is not increased. Experimental results show that the low pass filtering significantly improves the performance of the stochastic gradient algorithm and does not compromise the tracking of changes in the noise scenario. In addition, experiments demonstrate that the proposed stochastic gradient algorithm preserves the benefit of the SP-SDW-MWF over QIC-GSC. The limited computational cost and the better noise reduction performance of the proposed algorithm make it a good alternative to the SPA [14] for implementation in hearing aids.

Brief Description of the Drawings

A number of embodiments of the present invention, together with some aspects of the prior art will now be described with reference to the drawings, in which:

Fig. 1 depicts the concept of a Generalized Sidelobe Canceller;

Fig. 2 depicts an equivalent approach of multi-channel Wiener filtering;

Fig. 3 depicts a Spatially Pre-processed SDW MWF;

Fig. 4 depicts the decomposition of SP-SDW-MWF with w_0 in a multi-channel filter w_d and single-channel postfilter $e_1 - w_0$;

Fig. 5 shows the influence of $1/\mu$ on the performance of the SDR GSC for different gain mismatches Υ_2 at the second microphone;

Fig. 6 shows the influence of $1/\mu$ on the performance of the SP SDW MWF with w_0 for different gain mismatches Υ_2 at the second microphone;

Fig. 7 shows the $\Delta SNR_{intellig}$ and $SD_{intellig}$ for QIC-GSC as a function of β^2 for different gain mismatches Υ_2 at the second microphone;

Fig. 8 depicts the complexity of TD and FD Stochastic Gradient (SG) algorithm with LP filtering as a function of filter length L per channel; M = 3 (for comparison, the complexity of the standard NLMS ANC and SPA are depicted too);

Fig. 9 depicts the performance of different FD Stochastic Gradient (FD-SG) algorithms; (a) Stationary speechlike noise at 90°; (b) Multi-talker babble noise at 90°;

Fig. 10 depicts the influence of LP filter on performance of FD stochastic gradient SP-SDW-MWF ($1/\mu = 0.5$) without w_0 and with w_0. Babble noise at 90°;

Fig. 11 depicts the convergence behavior of FD-SG for $\lambda = 0$ and $\lambda = 0.9998$. The noise source position suddenly changes from 90° to 180° and vice versa;

Fig. 12 depicts the performance of FD stochastic gradient implementation of SP-SDW-MWF with LP ($\lambda = 0.9998$) in a multiple noise source scenario; and

Fig. 13 depicts the performance of FD SPA in a multiple noise source scenario.

Detailed Description

Before the invention is described in detail, the prior art GSC [4] and the QIC-GSC [14, 19] will be reviewed under section 1. Under section 2, the Multi-channel Wiener Filter (MWF) technique will be discussed [20].

1 Generalized Sidelobe Canceller (GSC)

1.1 Concept

Figure 1 describes the concept of the Generalized Sidelobe Canceller (GSC) [4], which consists of a fixed, spatial pre-processor, i.e., a fixed beamformer $\mathbf{A}(z)$ and a blocking matrix $\mathbf{B}(z)$, and an ANC. Given M microphone signals

$$u_i[k] = u_i^s[k] + u_i^n[k],\ i = 1,\ ...,\ M \tag{1}$$

with $u_i^s[k]$ the desired speech contribution and $u_i^n[k]$ the noise contribution, the fixed beamformer $\mathbf{A}(z)$ (e.g., delay-and-sum) creates a so-called speech reference

$$y_0[k] = y_0^s[k] + y_0^n[k], \tag{2}$$

by steering a beam towards the direction of the desired signal with a speech contribution $y_0^s[k]$ and a noise contribution $y_0^n[k]$. In the sequel an endfire array is assumed and the desired speaker is assumed to be in front at 0°. The blocking matrix $\mathbf{B}(z)$ creates $M-1$ so-called noise references

$$y_i[k] = y_i^s[k] + y_i^n[k],\ i = 1,\ ...,\ M-1 \tag{3}$$

by steering zeroes towards the front so that the noise contributions $y_i^n[k]$ are dominant compared to the speech leakage contributions $y_i^s[k]$. In the sequel, the superscripts s and n are used to refer to the speech and noise contribution of a signal. During periods of speech + noise, the references $y_i[k]$, $i = 0,\ ...,\ M-1$ contain speech + noise. During periods of noise only, $y_i[k]$, $i = 0,\ ...,\ M-1$ only consist of a noise component, i.e., $y_i[k] = y_i^n[k]$. The second order statistics of the noise signal are assumed to be quite stationary such that they can be estimated during periods of noise only.

To design the fixed, spatial pre-processor, assumptions are made about the microphone characteristics, the speaker position and the microphone positions and furthermore reverberation is assumed to be absent. If these assumptions are satisfied, the noise references do not contain any speech, i.e., $y_i^s[k] = 0$, for $i = 1,\ ...,\ M-1$. However, in practice, the assumptions are often violated (e.g. due to microphone mismatch and reverberation) so that speech leaks into the noise references. To limit the effect of such signal

leakage, the ANC $\mathbf{w}_{1:M-1}$[1]

$$\mathbf{w}_{1:M-1}^H = \begin{bmatrix} \mathbf{w}_1^H & \mathbf{w}_2^H & \dots & \mathbf{w}_{M-1}^H \end{bmatrix} \tag{4}$$

where

$$\mathbf{w}_i = \begin{bmatrix} w_i[0] & w_i[1] & \dots & w_i[L-1] \end{bmatrix}^T, \tag{5}$$

is adapted during periods of noise only [7, 13]. Hence, the ANC $\mathbf{w}_{1:M-1}$ minimizes the output noise power, i.e.,

$$\boxed{\mathbf{w}_{1:M-1} = \arg \min_{\mathbf{w}_{1:M-1}} \mathcal{E}\{|y_0^n[k-\Delta] - \mathbf{w}_{1:M-1}^H[k]\mathbf{y}_{1:M-1}^n[k]|^2\}} \tag{6}$$

and equals

$$\boxed{\mathbf{w}_{1:M-1} = \mathcal{E}\{\mathbf{y}_{1:M-1}^n \mathbf{y}_{1:M-1}^{n,H}\}^{-1} \mathcal{E}\{\mathbf{y}_{1:M-1}^n y_0^{n,*}[k-\Delta]\},} \tag{7}$$

where

$$\mathbf{y}_{1:M-1}^{n,H}[k] = \begin{bmatrix} \mathbf{y}_1^{n,H}[k] & \mathbf{y}_2^{n,H}[k] & \dots & \mathbf{y}_{M-1}^{n,H}[k] \end{bmatrix} \tag{8}$$

$$\mathbf{y}_i^n[k] = \begin{bmatrix} y_i^n[k] & y_i^n[k-1] & \dots & y_i^n[k-L+1] \end{bmatrix}^T \tag{9}$$

and where Δ is a delay applied to the speech reference to allow for non-causal taps in the filter $\mathbf{w}_{1:M-1}$. The delay Δ is usually set to $\lceil \frac{L}{2} \rceil$, where $\lceil x \rceil$ returns the smallest integer equal or larger than x. The subscript $1 : M-1$ in $\mathbf{w}_{1:M-1}$ and $\mathbf{y}_{1:M-1}$ refers to the subscripts of the first and last channel component of the adaptive filter and input vector, respectively.

Under ideal conditions ($y_i^s[k] = 0$, $i = 1, ,,,. M-1$), the GSC minimizes the residual noise while not distorting the desired speech signal, i.e., $z^s[k] = y_0^s[k-\Delta]$. However, when used in combination with small-sized arrays, a small error in the assumed signal model (hence $y_i^s[k] \neq 0$, $i = 1, ..., M-1$) already suffices to produce a significantly distorted output speech signal $z^s[k]$

$$z^s[k] = y_0^s[k-\Delta] - \mathbf{w}_{1:M-1}^H \mathbf{y}_{1:M-1}^s[k], \tag{10}$$

even when only adapting during noise-only periods, so a robustness constraint on $\mathbf{w}_{1:M-1}$ is required [17]. In addition, the fixed beamformer $\mathbf{A}(z)$ should be designed so that the distortion in the speech reference $y_0^s[k]$ is minimal for all possible model errors. In the sequel, a delay-and-sum beamformer is used. For small-sized arrays, this beamformer offers sufficient robustness against signal model errors, as it minimizes the white noise gain or noise sensitivity [2]. Given statistical knowledge about the signal model errors that occur in practice, further optimized beamformers can be designed, e.g., using the techniques in [31].

[1] In a time-domain implementation, the input signals of the adaptive filter $\mathbf{w}_{1:M-1}$ and the filter $\mathbf{w}_{1:M-1}$ are real. Hence, $\mathbf{w}_{1:M-1}^H = \mathbf{w}_{1:M-1}^T$. In the sequel, the formulas are generalized to complex input signals so that they can also be applied to a subband implementation.

[2] The white noise gain or noise sensitivity is defined as the ratio of the spatially white noise gain to the gain of the desired signal and is often used to quantify the sensitivity of an algorithm against errors in the assumed signal model [2, 14].

1.2 Quadratic Inequality Constraint (QIC-GSC)

A common approach to increase the robustness of the GSC is to apply a Quadratic Inequality Constraint (QIC) [9]-[14, 19] to the ANC filters $\mathbf{w}_{1:M-1}$, so that the optimization criterion (6) of the GSC is modified into

$$\begin{aligned} \mathbf{w}_{1:M-1} &= \arg\min_{\mathbf{w}_{1:M-1}} \mathcal{E}\{|y_0^n[k-\Delta] - \mathbf{w}_{1:M-1}^H[k]\mathbf{y}_{1:M-1}^n[k]|^2\} \\ &\text{subject to } \mathbf{w}_{1:M-1}^H\mathbf{w}_{1:M-1} \leq \beta^2. \end{aligned} \tag{11}$$

The QIC avoids excessive growth of the filter coefficients $\mathbf{w}$. Hence, it reduces the undesired speech distortion when speech leaks into the noise references. In [14, 19], it is shown that -for a GSC with a blocking matrix $\mathbf{B}(f)$ that satisfies $\mathbf{B}^H(f)\mathbf{B}(f) = \mathbf{I}$- the QIC on the ANC filters corresponds to a constraint on the noise sensitivity.

In [14], the QIC-GSC is implemented by using the adaptive *scaled projection algorithm*: at each update step, the quadratic constraint is applied to the newly obtained ANC filter by scaling the filter coefficients by $\frac{\beta}{\|\mathbf{w}_{1:M-1}\|}$ when $\mathbf{w}_{1:M-1}^H\mathbf{w}_{1:M-1}$ exceeds β^2. Although this technique works well for LMS updating, it does not appear to be as effective for RLS as for LMS [19]. Recently, Tian et al. implemented the quadratic constraint by using *variable loading* [19]. For RLS, this technique provides a better approximation to the optimal solution (11) than the scaled projection algorithm. For LMS, variable loading does not appear to offer any performance advantage over the cheaper, scaled projection LMS.

2 Multi-channel Wiener filtering (MWF)

2.1 Concept

Recently, a Multi-channel Wiener filtering (MWF) technique has been proposed that provides a Minimum Mean Square Error (MMSE) estimate of the desired signal portion in one of the received microphone signals [21, 22, 23, 24]. In contrast to the GSC, this filtering technique does not make any a priori assumptions about the signal model and is found to be more robust [16, 17, 21]. Especially in complicated noise scenarios such as multiple noise sources or diffuse noise, the MWF outperforms the GSC, even when the GSC is supplied with a robustness constraint [17].

The MWF $\bar{\mathbf{w}}_{1:M} \in C^{ML\times 1}$ minimizes the Mean Square Error (MSE) between a delayed version of the (unknown) speech signal $u_i^s[k-\Delta]$ at the i-th (e.g., first) microphone and the sum $\bar{\mathbf{w}}_{1:M}^H\mathbf{u}_{1:M}[k]$ of the M filtered, received microphone signals:

$$\bar{\mathbf{w}}_{1:M} = \arg\min_{\bar{\mathbf{w}}_{1:M}} \mathcal{E}\left\{\left|u_i^s[k-\Delta] - \bar{\mathbf{w}}_{1:M}^H\mathbf{u}_{1:M}[k]\right|^2\right\}, \tag{12}$$

leading to:

$$\boxed{\bar{\mathbf{w}}_{1:M} = \mathcal{E}\{\mathbf{u}_{1:M}[k]\mathbf{u}_{1:M}^H[k]\}^{-1}\mathcal{E}\{\mathbf{u}_{1:M}[k]u_i^{s,*}[k-\Delta]\},} \tag{13}$$

with

$$\bar{\mathbf{w}}_{1:M}^H = \left[\begin{array}{cccc} \bar{\mathbf{w}}_1 & \bar{\mathbf{w}}_2 & \cdots & \bar{\mathbf{w}}_M \end{array}\right]^T, \tag{14}$$

$$\mathbf{u}_{1:M}^H[k] = \left[\begin{array}{cccc} \mathbf{u}_1[k] & \mathbf{u}_2[k] & \cdots & \mathbf{u}_M[k] \end{array}\right]^H, \tag{15}$$

$$\mathbf{u}_i[k] = \left[\begin{array}{cccc} u_i[k] & u_i[k-1] & \cdots & u_i[k-L+1] \end{array}\right]^T. \tag{16}$$

An equivalent approach consists in estimating a delayed version of the (unknown) noise signal $u_i^n[k-\Delta]$ in the i-th microphone, resulting in

$$\boxed{\mathbf{w}_{1:M} = \arg\min_{\mathbf{w}_{1:M}} \mathcal{E}\left\{\left|u_i^n[k-\Delta] - \bar{\mathbf{w}}_{1:M}^H\mathbf{u}_{1:M}[k]\right|^2\right\},} \tag{17}$$

and

$$\boxed{\mathbf{w}_{1:M} = \mathcal{E}\{\mathbf{u}_{1:M}[k]\mathbf{u}_{1:M}^H[k]\}^{-1}\mathcal{E}\{\mathbf{u}_{1:M}[k]u_i^{n,*}[k-\Delta]\},} \tag{18}$$

where

$$\mathbf{w}_{1:M}^H = \left[\begin{array}{cccc} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_M \end{array}\right]^T. \tag{19}$$

The estimate of the speech component $u_i^s[k-\Delta]$ is then obtained by subtracting the estimate $\widehat{u_i^n}[k-\Delta] = \mathbf{w}_{1:M}^H\mathbf{u}_{1:M}[k]$ from the delayed, i-th microphone signal $u_i[k-\Delta]$, i.e.

$$\widehat{u_i^s}[k-\Delta] = u_i[k-\Delta] - \mathbf{w}_{1:M}^H\mathbf{u}_{1:M}[k]. \tag{20}$$

This is depicted in Figure 2 for $u_i^n[k-\Delta] = u_1^n[k-\Delta]$. Using (13) and (18), it can be easily shown that

$$\mathbf{w}_{1:M} + \bar{\mathbf{w}}_{1:M} = \mathbf{e}_{(i-1)L+\Delta}, \tag{21}$$

with $\mathbf{e}_l$ the l-th canonical vector, defined as

$$\mathbf{e}_l = \left[\begin{array}{ccccccc} 0 & \cdots & 0 & \underbrace{1}_{\text{position } l} & 0 & \cdots & 0 \end{array}\right]^T. \tag{22}$$

This shows that the two approaches indeed lead to exactly the same speech signal estimate. A procedure for computing $\mathbf{w}_{1:M}$ or $\bar{\mathbf{w}}_{1:M}$ will be given in Section 2.3.

2.2 Trade-off speech distortion versus noise reduction (SDW-MWF)

The residual error energy equals

$$\mathcal{E}\{|e[k]|^2\} = \mathcal{E}\{\left|u_i^s[k-\Delta] - \bar{\mathbf{w}}_{1:M}^H \mathbf{u}_{1:M}[k]\right|^2\}, \tag{23}$$

and can be decomposed as

$$\underbrace{\mathcal{E}\{\left|u_i^s[k-\Delta] - \bar{\mathbf{w}}_{1:M}^H \mathbf{u}_{1:M}^s[k]\right|^2\}}_{\epsilon_d^2} + \underbrace{\mathcal{E}\{\left|\bar{\mathbf{w}}_{1:M}^H \mathbf{u}_{1:M}^n[k]\right|^2\}}_{\epsilon_n^2} \tag{24}$$

where ϵ_d^2 equals the speech distortion energy and ϵ_n^2 the residual noise energy. The design criterion of the MWF can be generalized to allow for a trade-off between speech distortion and noise reduction, by incorporating a weighting factor μ [20] with $\mu \in [0, \infty]$

$$\boxed{\bar{\mathbf{w}}_{1:M} = \arg\min_{\bar{\mathbf{w}}_{1:M}} \mathcal{E}\{\left|u_i^s[k-\Delta] - \bar{\mathbf{w}}_{1:M}^H \mathbf{u}_{1:M}^s[k]\right|^2\} + \mu\mathcal{E}\{\left|\bar{\mathbf{w}}_{1:M}^H \mathbf{u}_{1:M}^n[k]\right|^2\}.} \tag{25}$$

The solution of (13) is given by

$$\boxed{\bar{\mathbf{w}}_{1:M} = \mathcal{E}\{\mathbf{u}_{1:M}^s[k]\mathbf{u}_{1:M}^{s,H}[k] + \mu\mathbf{u}_{1:M}^n[k]\mathbf{u}_{1:M}^{n,H}[k]\}^{-1}\mathcal{E}\{\mathbf{u}_{1:M}^s[k]u_i^{s,*}[k-\Delta]\},} \tag{26}$$

which corresponds to the Wiener formula with an adjustable input noise level. Note that (18) is obtained with $\mu = 1$ and that (21) still applies. The filter (26) corresponds to the time-domain constrained estimator proposed in [32], which optimizes the following criterion:

$$\min_{\bar{\mathbf{w}}} \epsilon_d^2 \text{ subject to } \epsilon_n^2 \leq \alpha\mathcal{E}\{\mathbf{u}_{1:M}^{n,H}\mathbf{u}_{1:M}^n\} \tag{27}$$

where $0 \leq \alpha \leq 1$ and μ is the Lagrange-multiplier.

Equivalently, the optimization criterion for $\bar{\mathbf{w}}$ in (13) can be modified into

$$\boxed{\mathbf{w}_{1:M} = \arg\min_{\mathbf{w}_{1:M}} \mathcal{E}\{\left|\mathbf{w}_{1:M}^H \mathbf{u}_{1:M}^s[k]\right|^2\} + \mu\mathcal{E}\{\left|u_i^n[k-\Delta] - \mathbf{w}_{1:M}^H \mathbf{u}_{1:M}^n[k]\right|^2\},} \tag{28}$$

resulting in

$$\boxed{\mathbf{w}_{1:M} = \mathcal{E}\{\mathbf{u}_{1:M}^n[k]\mathbf{u}_{1:M}^{n,H}[k] + \frac{1}{\mu}\mathbf{u}_{1:M}^s[k]\mathbf{u}_{1:M}^{s,H}[k]\}^{-1}\mathcal{E}\{\mathbf{u}_{1:M}^n[k]u_i^{n,*}[k-\Delta]\}.} \tag{29}$$

In the sequel, we will refer to (29) as the *Speech Distortion Weighted* Multi-channel Wiener Filter (SDW-MWF).

The factor $\mu \in [0, \infty]$ trades off speech distortion versus noise reduction. If $\mu = 1$, the MMSE criterion

(12) or (17) is obtained. If $\mu > 1$, the residual noise level will be reduced at the expense of increased speech distortion. By setting μ to ∞, all emphasis is put on noise reduction and speech distortion is completely ignored. This results in $\bar{\mathbf{w}} = 0$ or $\mathbf{w} = \mathbf{e}_{(i-1)L+\Delta}$, which means that the output signal equals 0. Setting μ to 0 on the other hand, results in $\bar{\mathbf{w}} = \mathbf{e}_{(i-1)L+\Delta}$ or $\mathbf{w} = 0$ and hence in no noise reduction.

2.3 Implementation of MWF

In practice, the correlation matrix $\mathcal{E}\{\mathbf{u}^s_{1:M}[k]\mathbf{u}^{s,H}_{1:M}[k]\}$ is unknown. During periods of speech, the inputs $u_i[k]$ consist of speech + noise, i.e., $u_i[k] = u^s_i[k] + u^n_i[k]$, $i = 1, ..., M$. During periods of noise, only the noise component $u^n_i[k]$ is observed. Assuming that the speech and noise signal are uncorrelated, $\mathcal{E}\{\mathbf{u}^s_{1:M}[k]\mathbf{u}^{s,H}_{1:M}[k]\}$ can be estimated as

$$\mathcal{E}\{\mathbf{u}^s_{1:M}[k]\mathbf{u}^{s,H}_{1:M}[k]\} = \mathcal{E}\{\mathbf{u}_{1:M}[k]\mathbf{u}^H_{1:M}[k]\} - \mathcal{E}\{\mathbf{u}^n_{1:M}[k]\mathbf{u}^{n,H}_{1:M}[k]\}, \tag{30}$$

where the second order statistics $\mathcal{E}\{\mathbf{u}_{1:M}[k]\mathbf{u}^H_{1:M}[k]\}$ are estimated during speech + noise and the statistics $\mathcal{E}\{\mathbf{u}^n_{1:M}[k]\mathbf{u}^{n,H}_{1:M}[k]\}$ during periods of noise only. Like for the GSC, a robust speech detection is thus needed. Using (30), (29) and (26) can be re-written as:

$$\boxed{\mathbf{w}_{1:M} = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{u}_{1:M}[k]\mathbf{u}^H_{1:M}[k]\} + (1-\frac{1}{\mu})\mathcal{E}\{\mathbf{u}^n_{1:M}[k]\mathbf{u}^{n,H}_{1:M}[k]\}\right)^{-1} \mathcal{E}\{\mathbf{u}^n_{1:M}[k]u^{n,*}_i[k-\Delta]\}} \tag{31}$$

and

$$\boxed{\begin{aligned}\bar{\mathbf{w}}_{1:M} &= \left(\mathcal{E}\{\mathbf{u}_{1:M}[k]\mathbf{u}^H_{1:M}[k]\} + (\mu-1)\mathcal{E}\{\mathbf{u}^n_{1:M}[k]\mathbf{u}^{n,H}_{1:M}[k]\}\right)^{-1} \\ &\quad \times \left(\mathcal{E}\{\mathbf{u}_{1:M}[k]u^*_i[k-\Delta]\} - \mathcal{E}\{\mathbf{u}^n_{1:M}[k]u^{n,*}_i[k-\Delta]\}\right).\end{aligned}} \tag{32}$$

In [21], the Wiener filter is computed at each time instant k by means of a Generalized Singular Value Decomposition (GSVD) of an speech + noise and noise data matrix. A cheaper recursive alternative based on a QR-decomposition has been proposed in [22]. In [23, 24], a subband implementation has been developed to increase intelligibility and reduce complexity, making it suitable for hearing aid applications.

Finally note that instead of estimating $\mathcal{E}\{\mathbf{u}^s_{1:M}[k]\mathbf{u}^{s,H}_{1:M}[k]\}$ online using (30), a pre-determined estimate of $\mathcal{E}\{\mathbf{u}^s_{1:M}[k]\mathbf{u}^{s,H}_{1:M}[k]\}$ is sometimes used [25, 33]. In [25], this estimate is derived from clean speech recordings measured during an initial calibration phase. Additional recordings of the source speech signal allow to produce an estimate of the non-reverberant source speech signal instead of an estimate of the reverberant speech component in one of the microphone signals. However, since the room acoustics, the position of desired speaker and microphone characteristics may change over time, frequent re-calibration is required. In [33], a mathematical estimate of the correlation matrix and the correlation vector of the non-reverberant speech is exploited in which some signal model errors are taken into account.

In this Section, the present invention is described in detail.

In Section 3, the proposed adaptive multi-channel noise reduction technique, referred to as Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener filter, is described.

Section 3.2 describes a *first embodiment,* referred to as *Speech Distortion Regularized GSC* (SDR-GSC). A new design criterion is developed for the adaptive stage of the GSC: the ANC design criterion is supplemented with a regularization term that limits speech distortion due to signal model errors. In the SDR-GSC, a parameter μ is incorporated that allows for a trade-off between speech distortion and noise reduction. Focussing all attention to noise reduction, results in the standard GSC, while, on the other hand, focussing all attention towards speech distortion results in the output of the fixed beamformer. In noise scenarios with low SNR, adaptivity in the SDR-GSC can be easily reduced or excluded by increasing attention towards speech distortion, i.e., by decreasing the parameter μ to 0. The SDR-GSC is an alternative technique to the QIC-GSC to decrease the sensitivity of the GSC to signal model errors such as microphone mismatch, reverberation, ... In contrast to the QIC-GSC, the SDR-GSC shifts emphasis towards speech distortion when the amount of speech leakage grows. In the absence of signal model errors, the performance of the GSC is preserved. As a result, a better noise reduction performance is obtained for small model errors, while guaranteeing robustness against large model errors.

In a *second embodiment,* described in Section 3.3, we further improve the noise reduction performance of the SDR-GSC by adding an extra adaptive filtering operation w_0 on the speech reference signal. We refer to this generalized scheme as *Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filter* (SP-SDW-MWF). The SP-SDW-MWF is depicted in Figure 3 and encompasses the MWF as a special case. Again, a parameter μ is incorporated in the design criterion to allow for a trade-off between speech distortion and noise reduction. Focussing all attention to speech distortion, results in the output of the fixed beamformer.. Also here, adaptivity can be easily reduced or excluded by decreasing μ to 0. It is shown that -in the absence of speech leakage and for infinitely long filter lengths- the SP-SDW-MWF corresponds to a cascade of a SDR-GSC with a SDW-SWF postfilter. In the presence of speech leakage, the SP-SDW-MWF with w_0 tries to preserve its performance: compared to a SDR-GSC with SDW-SWF postfilter, the SP-SDW-MWF then contains extra filtering operations that compensate for the performance degradation of the SDR-GSC with SDW-SWF due to speech leakage. In contrast to the SDR-GSC (and thus also the GSC), performance does not degrade due to microphone mismatch. In [22, 27] recursive implementations of the (SDW-)MWF have been proposed based on a GSVD or QR decomposition. A subband implementation [28] results in improved intelligibility at a significantly lower cost compared to the fullband approach. These techniques[3] can be extended to implement the SDR-GSC and, more generally, the SP-SDW-MWF.

In a *third embodiment*, described in Section 4, we propose cheap *time-domain and frequency-domain stochastic gradient implementations* of the SDR-GSC and SP-SDW-MWF. Starting from the design criterion of the SDR-GSC, or more generally, the SP-SDW-MWF, we derive a time-domain stochastic gradient

[3]The implementation based on GSVD can only be used for the SP-SDW-MWF with filter w_0.

algorithm. In addition, we modify the LMS based algorithm [26] so that it applies to the SP-SDW-MWF. To increase convergence and reduce complexity, a frequency-domain implementation has been proposed. Both, the stochastic gradient and LMS based algorithm suffer from a large excess error when applied in highly time-varying noise scenarios. We show that the excess error in the stochastic gradient algorithm is reduced by applying a low pass filter to the part of the gradient estimate that limits speech distortion. The low pass filtering avoids a highly time-varying distortion of the desired speech component while not degrading the tracking performance needed in time-varying noise scenarios. The stochastic gradient SP-SDW-MWF outperforms the LMS based algorithm, while complexity is not increased. Experimental results show that the low pass filtering significantly improves the performance of the stochastic gradient algorithm and does not compromise the tracking of changes in the noise scenario. In addition, experiments demonstrate that the proposed stochastic gradient algorithm preserves the benefit of the SP-SDW-MWF over QIC-GSC. The limited computational cost and the better noise reduction performance of the proposed algorithm make it a good alternative to the SPA [14] for implementation in hearing aids.

3 Spatially pre-processed SDW Multi-channel Wiener filter

3.1 Concept

Figure 3 describes the Spatially pre-processed, Speech Distortion Weighted Multi-channel Wiener filter (SP-SDW-MWF). The SP-SDW-MWF consists of a fixed, spatial pre-processor, i.e., a fixed beamformer $\mathbf{A}(z)$ and a blocking matrix $\mathbf{B}(z)$, and an adaptive Speech Distortion Weighted Multi-channel Wiener filter (SDW-MWF). Given M microphone signals

$$u_i[k] = u_i^s[k] + u_i^n[k],\ i = 1,\ ...,\ M \tag{33}$$

with $u_i^s[k]$ the desired speech contribution and $u_i^n[k]$ the noise contribution, the fixed beamformer $\mathbf{A}(z)$ creates a so-called speech reference

$$y_0[k] = y_0^s k] + y_0^n[k], \tag{34}$$

by steering a beam towards the direction of the desired signal with a speech contribution $y_0^s[k]$ and a noise contribution $y_0^n[k]$. In the sequel an endfire array is assumed and the desired speaker is assumed to be in front at 0°. To preserve the robustness advantage of the MWF, the fixed beamformer $\mathbf{A}(z)$ should be designed so that the distortion in the speech reference $y_0^s[k]$ is minimal for all possible errors in the assumed signal model such as microphone mismatch. In the sequel, a delay-and-sum beamformer is used. For small-sized arrays, this beamformer offers sufficient robustness against signal model errors as it minimizes the white noise gain or noise sensitivity[4]. Given statistical knowledge about the signal model errors that occur in practice, a further optimized beamformer $\mathbf{A}(z)$ can be designed, e.g., using the techniques in [31]. The

[4]The white noise gain or noise sensitivity is defined as the ratio of the spatially white noise gain to the gain of the desired signal and is often used to quantify the sensitivity of an algorithm against errors in the assumed signal model [2, 14].

blocking matrix $\mathbf{B}(z)$ creates $M-1$ so-called noise references

$$y_i[k] = y_i^s[k] + y_i^n[k],\ i = 1,\ ...,\ M-1 \tag{35}$$

by steering zeroes towards the front so that the noise contributions $y_i^n[k]$ are dominant compared to the speech leakage contributions $y_i^s[k]$. A simple technique to create the noise references consists of pairwise subtracting the for 0° time-aligned microphone signals. Using [31, 34], further optimized noise references can be created. Speech leakage can then be minimized for a specified angular region around 0° instead of for 0°only, e.g., for an angular region from −20° to 20°. In addition, given statistical knowledge about the signal model errors that occur in practice, speech leakage can be minimized for all possible model errors by using [31].

In the sequel, the superscripts s and n are used to refer to the speech and noise contribution of a signal. During periods of speech + noise, the references $y_i[k]$, $i = 0,\ ...,\ M-1$ contain speech + noise. During periods of noise only, $y_i[k]$, $i = 0,\ ...,\ M-1$ only consist of a noise component, i.e., $y_i[k] = y_i^n[k]$. The second order statistics of the noise signal are assumed to be quite stationary such that they can be estimated during periods of noise only.

The SDW-MWF filter[5] $\mathbf{w}_{0:M-1}$

$$\boxed{\mathbf{w}_{0:M-1} = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s_{0:M-1}\mathbf{y}^{s,H}_{0:M-1}[k]\} + \mathcal{E}\{\mathbf{y}^n_{0:M-1}\mathbf{y}^{n,H}_{0:M-1}[k]\}\right)^{-1}\mathcal{E}\{\mathbf{y}^n_{0:M-1}y_0^{n,*}[k-\Delta]\},} \tag{36}$$

with

$$\mathbf{w}^H_{0:M-1}[k] = \left[\ \mathbf{w}_0^H[k]\ \ \mathbf{w}_1^H[k]\ \ ...\ \ \mathbf{w}_{M-1}^H[k]\ \right], \tag{37}$$

$$\mathbf{w}_i[k] = \left[\ w[0]\ \ w[1]\ \ ...\ \ w[L-1]\ \right]^T \tag{38}$$

$$\mathbf{y}^H_{0:M-1}[k] = \left[\ \mathbf{y}_0^H[k]\ \ \mathbf{y}_1^H[k]\ \ ...\ \ \mathbf{y}_{M-1}^H[k]\ \right], \tag{39}$$

$$\mathbf{y}_i[k] = \left[\ y_i[k]\ \ y_i[k-1]\ \ ...\ \ y_i[k-L+1]\ \right]^T, \tag{40}$$

provides an estimate $\mathbf{w}^H_{0:M-1}\mathbf{y}_{0:M-1}k]$ of the noise contribution $\mathbf{y}_0^n[k-\Delta]$[6] in the speech reference by minimizing the cost function $J(\mathbf{w}_{0:M-1})$

$$\boxed{J(\mathbf{w}_{0:M-1}) = \frac{1}{\mu}\underbrace{\mathcal{E}\{|\mathbf{w}^H_{0:M-1}[k]\mathbf{y}^s_{0:M-1}[k]|^2\}}_{\varepsilon_d^2} + \underbrace{\mathcal{E}\{|y_0^n[k-\Delta] - \mathbf{w}^H_{0:M-1}[k]\mathbf{y}^n_{0:M-1}[k]|^2\}}_{\varepsilon_n^2}.} \tag{41}$$

[5] In a time-domain implementation, the input signals of the adaptive filter and the filter $\mathbf{w}_{0:M-1}$ are real and hence, $\mathbf{w}^H_{0:M-1} = \mathbf{w}^T_{0:M-1}$. In the sequel, the formulas are generalized to complex input signals so that they can also be applied to a subband implementation.

[6] The delay Δ is applied to the speech reference to make the filter w non-causal. Usually, it is set to $\lceil \frac{L}{2} \rceil$, where $\lceil x \rceil$ returns the smallest integer equal or larger than x.

The subscript $0 : M-1$ in $\mathbf{w}_{0:M-1}$ and $\mathbf{y}_{0:M-1}$ refers to the subscripts of the first and last channel component of the adaptive filter and input vector, respectively. The term ε_d^2 represents the speech distortion energy and ε_n^2 the residual noise energy. The term $\frac{1}{\mu}\varepsilon_d^2$ in the cost function (41) limits the possible amount of speech distortion at the output of the SP-SDW-MWF. Hence, the SP-SDW-MWF adds robustness against signal model errors to the GSC by taking speech distortion explicitly into account in the design criterion of the adaptive stage. The parameter $\frac{1}{\mu} \in [0, \infty)$ trades off between noise reduction and speech distortion: the larger $\frac{1}{\mu}$, the smaller the amount of possible speech distortion. For $\mu = 0$, the output of the fixed beamformer $\mathbf{A}(z)$, delayed by Δ samples is obtained. In noise scenarios with very low Signal-to-Noise Ratio (SNR), e.g., -10 dB, a fixed beamformer may be preferred. Adaptivity can be easily reduced or excluded in the SP-SDW-MWF by decreasing μ to 0. Alternatively, adaptivity can be limited by applying a QIC to $\mathbf{w}_{0:M-1}$.

Note that when the fixed beamformer $\mathbf{A}(z)$ and the blocking matrix $\mathbf{B}(z)$ are set to

$$\mathbf{A}(z) = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^H \tag{42}$$

$$\mathbf{B}(z) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix}^H, \tag{43}$$

we obtain the original SDW-MWF that operates on the received microphone signals $u_i[k]$, $i = 1, \dots, M$.

Below, the different parameter settings of the SP-SDW-MWF are discussed. Depending on the setting of the parameter μ and presence or absence of the filter $\mathbf{w}_0$, the GSC, the (SDW-)MWF as well as in-between solutions such as the Speech Distortion Regularized GSC (SDR-GSC) may be obtained. We distinguish between two cases, i.e., the case where no filter $\mathbf{w}_0$ is applied to the speech reference (filter length $L_0 = 0$) and the case where an additional filter $\mathbf{w}_0$ is used ($L_0 \neq 0$).

The adaptive stage of the SP-SDW-MWF can be implemented using the recursive QRD-based implementation of the SDW-MWF [22]. Like for the SDW-MWF, complexity can be reduced by a subband implementation [23]. For $L_0 \neq 0$, also the GSVD based algorithm [20] can be applied. Cheaper stochastic gradient based algorithms are proposed in Section 4.

3.2 First embodiment: SDR-GSC, i.e., SP-SDW-MWF without $\mathbf{w}_0$

First, consider *the case without* $\mathbf{w}_0$, i.e. $L_0 = 0$. The solution for $\mathbf{w}_{1:M-1}$ in (36) then reduces to

$$\boxed{\arg \min_{\mathbf{w}_{1:M-1}} \frac{1}{\mu} \underbrace{\mathcal{E}\{|\mathbf{w}_{1:M-1}^H \mathbf{y}_{1:M-1}^s[k]|^2\}}_{\varepsilon_d^2} + \underbrace{\mathcal{E}\{|y_0^n[k-\Delta] - \mathbf{w}_{1:M-1}^H \mathbf{y}_{1:M-1}^n[k]|^2\}}_{\varepsilon_n^2},} \tag{44}$$

leading to

$$\boxed{\mathbf{w}_{1:M-1} = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s_{1:M-1}\mathbf{y}^{s,H}_{1:M-1}[k]\} + \mathcal{E}\{\mathbf{y}^n_{1:M-1}\mathbf{y}^{n,H}_{1:M-1}[k]\}\right)^{-1} \mathcal{E}\{\mathbf{y}^n_{1:M-1}[k]y_0^{n,*}[k-\Delta]\},} \quad (45)$$

where ε_d^2 is the speech distortion energy and ε_n^2 the residual noise energy.

Remark: For $L_0 = 0$, it is readily seen that (21) does not hold, i.e., $\mathbf{w}_{1:M-1} + \bar{\mathbf{w}}_{1:M-1} \neq \mathbf{e}_\Delta$ where

$$\bar{\mathbf{w}}_{1:M-1} = \left(\mathcal{E}\{\mathbf{y}^s_{1:M-1}\mathbf{y}^{s,H}_{1:M-1}\} + \mu\mathcal{E}\{\mathbf{y}^n_{1:M-1}\mathbf{y}^{n,H}_{1:M-1}\}\right)^{-1} \mathcal{E}\{\mathbf{y}^{s,H}_{1:M-1}y_0^{s,*}[k-\Delta]\}, \quad (46)$$

because the speech component $\mathbf{y}^s_{1:M-1}[k]$ in the input to the adaptive filter $\mathbf{w}_{1:M-1}$does not contain the estimated speech signal $y_0^s[k-\Delta]$.

If $\mu = 1$, the classical MMSE criterion (cfr. (17)) is obtained.

Compared to the optimization criterion (6) of the GSC, a regularization term

$$\frac{1}{\mu}\mathcal{E}\{\left|\mathbf{w}^H_{1:M-1}\mathbf{y}^s_{1:M-1}[k]\right|^2\} \quad (47)$$

has been added. This regularization term limits the amount of speech distortion that is caused by the filter $\mathbf{w}_{1:M-1}$ when speech leaks into the noise references, i.e., $y_i^s[k] \neq 0$, $i = 1, ..., M-1$. In the sequel, we therefore refer to the SP-SDW-MWF with $L_0 = 0$ as *Speech Distortion Regularized GSC (SDR-GSC)*. The smaller μ, the smaller the resulting amount of speech distortion will be. For $\mu = 0$, the output of the fixed beamformer $\mathbf{A}(z)$ delayed by Δ samples, is obtained. For $\mu = \infty$, all emphasis is put on noise reduction and speech distortion is not taken into account. This corresponds to the GSC. Hence, the SDR-GSC encompasses the GSC as a special case.

The regularization term $\frac{1}{\mu}\mathcal{E}\{\left|\mathbf{w}^H_{1:M-1}[k]\mathbf{y}^s_{1:M-1}[k]\right|^2\}$ with $\frac{1}{\mu} \neq 0$ adds robustness to the GSC, while not affecting the noise reduction performance in the absence of speech leakage.

- In the *absence of speech leakage, i.e.,* $y_i^s[k] = 0$, $i = 1, ..., M-1$, the regularization term equals 0 for all $\mathbf{w}_{1:M-1}$ and hence the residual noise energy ε_n^2 is effectively minimized. In other words, in the absence of speech leakage, the GSC solution is obtained.

- In the *presence of speech leakage,* i.e., $y_i^s[k] \neq 0$, $i = 1, ..., M-1$, speech distortion is taken into account in the optimization criterion (44) for the adaptive filter $\mathbf{w}$, limiting speech distortion plus reducing noise. The larger the amount of speech leakage, the more attention is paid to speech distortion.

  To limit speech distortion alternatively, a QIC is often imposed on the filter $\mathbf{w}_{1:M-1}$ (see Section 1.2). In contrast to the SDR-GSC, the QIC acts irrespective of the amount of speech leakage $\mathbf{y}^s[k]$ that is present. The constraint value β^2 in (11) has to be chosen based on the largest model errors that may occur. As a consequence, noise reduction performance is compromised even when no or very small

model errors are present. Hence, the QIC is more conservative than the SDR-GSC. The experimental results in Section 3.4 confirm this.

3.3 Second embodiment: SP-SDW-MWF with filter $\mathbf{w}_0$

Since the SDW-MWF (36) takes speech distortion explicitly into account in its optimization criterion, an additional filtering $\mathbf{w}_0$ on the speech reference $y_0[k]$ may be added. The SDW-MWF (36) then solves the following more general optimization criterion

$$\mathbf{w}_{0:M-1} = \arg\min_{\mathbf{w}_{0:M-1}} \underbrace{\mathcal{E}\left\{\left|y_0^n[k-\Delta] - \begin{bmatrix} \mathbf{w}_0^H & \mathbf{w}_{1:M-1}^H \end{bmatrix} \begin{bmatrix} \mathbf{y}_0^n[k] \\ \mathbf{y}_{1:M-1}^n[k] \end{bmatrix}\right|^2\right\}}_{\epsilon_n^2} + \frac{1}{\mu} \underbrace{\mathcal{E}\left\{\left|\begin{bmatrix} \mathbf{w}_0^H & \mathbf{w}_{1:M-1}^H \end{bmatrix} \begin{bmatrix} \mathbf{y}_0^s[k] \\ \mathbf{y}_{1:M-1}^s[k] \end{bmatrix}\right|^2\right\}}_{\epsilon_d^2}, \tag{48}$$

where $\mathbf{w}_{0:M-1}^H = [\mathbf{w}_0^H \ \mathbf{w}_{1:M-1}^H]$ is given by (36).

Again, μ trades off speech distortion and noise reduction. For $\mu = \infty$, speech distortion ϵ_d^2 is completely ignored so that the solution becomes

$$\begin{aligned} \mathbf{w}_{0:M-1}^H &= \begin{bmatrix} \mathbf{w}_0^H & \mathbf{w}_{1:M-1}^H \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{e}_\Delta^H & \mathbf{0}^H \end{bmatrix}, \end{aligned} \tag{49}$$

which results in a zero output signal. For $\mu = 0$, all attention is paid to speech distortion so that the output of the fixed beamformer delayed by Δ samples, is obtained.

- In the *absence of speech leakage*, i.e., $y_i^s[k] = 0$ for $i = 1, \ldots, M-1$, and for infinitely long filters $\mathbf{w}_i$, $i = 0, \ldots, M-1$, the SP-SDW-MWF with $\mathbf{w}_0$ corresponds to the cascade of a SDR-GSC and a SDW Single-channel WF (SDW-SWF) postfilter [30, 35].

  Proof: In case of infinite filter lengths, the SP-SDW-MWF $\mathbf{W}_{0:M-1}(f)$ and its optimization criterion can be represented in the frequency-domain:

$$\begin{aligned} \mathbf{W}_{0:M-1}(f) = \arg\min_{\mathbf{W}_{0:M-1}} & \mathcal{E}\left\{\left|\begin{bmatrix}(\exp(-j2\pi f\Delta) - W_0^*(f)) & -\mathbf{W}_{1:M-1}^H(f)\end{bmatrix}\begin{bmatrix} Y_0^n(f) \\ \mathbf{Y}_{1:M-1}^n(f)\end{bmatrix}\right|^2\right\} \\ & + \frac{1}{\mu}\mathcal{E}\left\{\left|\begin{bmatrix} W_0^*(f) & \mathbf{W}_{1:M-1}^H(f) \end{bmatrix}\begin{bmatrix} Y_0^s(f) \\ \mathbf{Y}_{1:M-1}^s(f)\end{bmatrix}\right|^2\right\} \end{aligned} \tag{50}$$

Without loss of generality, we assume -for reasons of simplicity- $\Delta = 0$.
Decompose $\mathbf{W}_{1:M-1}(f)$ as

$$\mathbf{W}_{1:M-1}(f) = (1 - W_0(f))\,\mathbf{W}_{d,1:M-1}(f) \tag{51}$$

with $W_0(f)$ a single-channel and $\mathbf{W}_{d,1:M-1}(f)$ a multi-channel filter and define an intermediate output $V(f)$ (see also Figure 4) as

$$V(f) = Y_0(f) - \mathbf{W}_{d,1:M-1}^H(f)\mathbf{Y}_{1:M-1}(f). \tag{52}$$

Then, the cost function $J(W_0, \mathbf{W}_{d,1:M-1})$ of (50) can be re-written as

$$J = \mathcal{E}\left\{|(1 - W_0^*(f))\,V^n(f)|^2\right\} + \frac{1}{\mu}\mathcal{E}\left\{|W_0^*(f)V^s(f) + \mathbf{W}_{d,1:M-1}^H(f)\mathbf{Y}_{1:M-1}^s(f)|^2\right\}. \tag{53}$$

From $\frac{\partial}{\partial W_0^*}J(W_0, \mathbf{W}_{d,1:M-1}) = 0$, we find

$$W_0(f) = \left(\mathcal{E}\{V^nV^{n,*}\} + \frac{1}{\mu}\mathcal{E}\{V^sV^{s,*}\}\right)^{-1}\left(\mathcal{E}\{V^nV^{n,*}\} - \frac{1}{\mu}\mathcal{E}\{V^s\mathbf{Y}_{1:M-1}^{s,H}\mathbf{W}_{d,1:M-1}\}\right), \tag{54}$$

This *single-channel filter* $W_0(f)$ consists of two terms.

- The first term

$$W_{0,1}(f) = \left(\mathcal{E}\{V^nV^{n,*}\} + \frac{1}{\mu}\mathcal{E}\{V^sV^{s,*}\}\right)^{-1}\mathcal{E}\{V^nV^{n,*}\} \tag{55}$$

  estimates the noise component $V^n(f)$ in the intermediate output $V(f)$. The filter $1 - W_{0,1}$ corresponds to a SDW Single-channel Wiener Filter (SDW-SWF) that estimates the speech component $V^s(f)$.

- The second term

$$W_{0,2}(f) = \left(\mathcal{E}\{V^nV^{n,*}\} + \frac{1}{\mu}\mathcal{E}\{V^sV^{s,*}\}\right)^{-1}\left(-\frac{1}{\mu}\mathcal{E}\{V^s\mathbf{Y}_{1:M-1}^{s,H}\mathbf{W}_{d,1:M-1}\}\right) \tag{56}$$

  estimates the speech leakage filtered by $\mathbf{W}_{d,1:M-1}(f)$, i.e., $-\mathbf{W}_{d,1:M-1}^H\mathbf{Y}_{1:M-1}^s$. The speech component in the intermediate output $V(f)$ equals $V^s(f) = Y_0^s - \mathbf{W}_{d,1:M-1}^H\mathbf{Y}_{1:M-1}^s$. The filter $W_{0,2}(f)$ tries to compensate for the distortion $-\mathbf{W}_{d,1:M-1}^H\mathbf{Y}_{1:M-1}^s$ by adding an estimate of $\mathbf{W}_{d,1:M-1}^H\mathbf{Y}_{1:M-1}^s$ to the output of the SDW-SWF.

In the absence of speech leakage (i.e., $\mathbf{Y}_{1:M-1}^s = \mathbf{0}$), the filter $W_{0,2}(f)$ equals zero and $1 - W_0(f)$ corresponds to a SDW-SWF.

From $\frac{\partial}{\partial \mathbf{W}_{d,1:M-1}} J(W_0, \mathbf{W}_{d,1:M-1}) = 0$, we obtain the following solution for $\mathbf{W}_{d,1:M-1}(f)$:

$$\mathbf{W}_{d,1:M-1}(f) = \left(\mathcal{E}\{\mathbf{Y}^n_{1:M-1}\mathbf{Y}^{n,H}_{1:M-1}\} + \frac{1}{\mu}\mathcal{E}\{\mathbf{Y}^s_{1:M-1}\mathbf{Y}^{s,H}_{1:M-1}\}\right)^{-1} \left(\mathcal{E}\{\mathbf{Y}^n_{1:M-1}Y^{n,*}_0\} - \frac{1}{\mu}\mathcal{E}\{\mathbf{Y}^s_{1:M-1}Y^{s,*}_0\frac{W_0}{1-W_0}\}\right). \quad (57)$$

Also the *multi-channel filter* $\mathbf{W}_{d,1:M-1}(f)$ consists of two terms.

- The first term corresponds to the SDR GSC

$$\left(\mathcal{E}\{\mathbf{Y}^n_{1:M-1}\mathbf{Y}^{n,H}_{1:M-1}\} + \frac{1}{\mu}\mathcal{E}\{\mathbf{Y}^s_{1:M-1}\mathbf{Y}^{s,H}_{1:M-1}\}\right)^{-1} \mathcal{E}\{\mathbf{Y}^n_{1:M-1}Y^{n,*}_0\} \quad (58)$$

  and estimates the noise component $Y^n_0(f)$ at the output of the fixed beamformer.
- The second term tries to compensate for the speech distortion $-W^*_0(f)Y^s_0(f)$ caused by $W_0(f)$ by adding an estimate of $\frac{W^*_0(f)}{1-W^*_0(f)}Y^s_0(f)$ to the output of the SDR-GSC. Note that this corresponds to adding an estimate of $W^*_0(f)Y^s_0(f)$ to the output $Z(f)$ of the SP-SDW-MWF.

In the absence of speech leakage, $\mathbf{W}_{d,1:M-1}(f)$ corresponds to a SDR-GSC or a GSC.
Figure 4 illustrates graphically the solution for $\mathbf{W}_{d,1:M-1}(f)$ and $W_0(f)$ for $\Delta = 0$. In the absence of speech leakage, the filters that try to compensate for the speech distortion equal 0, hence, the SP-SDW-MWF corresponds to a SDR-GSC (or GSC) with SDW-SWF postfilter. The SP-SDW-MWF achieves the same or a better Signal-to-Noise Ratio (SNR) improvement than the SDR-GSC, depending on the noise scenario. ■

3.4 Experimental results

This Section illustrates the theoretical results of Section 3.2 and Section 3.3 by means of experimental results for a hearing aid application. Section 3.4.1 and Section 3.4.2, respectively, describe the set-up and the performance measures that are used. In Section 3.4.3, the impact of the different parameter settings of the SP-SDW-MWF on the performance and the sensitivity to signal model errors is evaluated. Comparison is made with the QIC-GSC.

3.4.1 Set-up

A three-microphone Behind-The-Ear (BTE) hearing aid with three omnidirectional microphones (Knowles FG-3452) has been mounted on a dummy head in an office room. The interspacing d between the first and the second microphone is about $d = 1$ cm and the interspacing between the second and third microphone about 1.5 cm. The reverberation time $T_{60\,\mathrm{dB}}$ is about 700 ms for a speech weighted noise. The desired speech signal and the noise signals are uncorrelated. Both the speech and the noise signal have a level of

70 dB SPL at the center of the head. The desired speech source and noise sources are positioned at a distance of 1 meter from the head: the speech source in front of the head, the noise sources at an angle θ w.r.t. the speech source. To get an idea of the average performance based on directivity only, stationary speech and noise signals with the same, average long-term power spectral density are used. The signals can be found on [36]. The total duration of the input signal is 10 seconds of which 5 seconds contains noise only and 5 seconds contain both the speech and noise signal. For evaluation purposes, the speech and noise signal have been recorded separately.

The microphone signals are pre-whitened prior to processing to improve intelligibility [37], and the output is accordingly de-whitened. In the experiments, the microphones have been calibrated by means of recordings of an anechoic speech weighted noise signal positioned at 0° measured while the microphone array was mounted on the head. A delay-and-sum beamformer is used as a fixed beamformer, since -in case of small microphone interspacing - it is robust to model errors. The blocking matrix $\mathbf{B}$ pairwise subtracts the time aligned calibrated microphone signals.

To investigate the effect of the different parameter settings (i.e. μ, $\mathbf{w}_0$) on the performance only, the filter coefficients are computed using (36) where $\mathcal{E}\{\mathbf{y}^s_{0:M-1}\mathbf{y}^{s,H}_{0:M-1}\}$ is estimated by means of the clean speech contributions of the microphone signals. In practice, $\mathcal{E}\{\mathbf{y}^s_{0:M-1}\mathbf{y}^{s,H}_{0:M-1}\}$ is approximated using (30). The effect of approximation (30) on the performance was found to be small (i.e. differences of at most 0.5 dB in intelligibility weighted Signal-to-Noise ratio improvement) for the given data set. The QIC-GSC is implemented using variable loading RLS [19]. The filter length L per channel equals 96.

3.4.2 Performance measures

To assess the performance of the different approaches, the broadband intelligibility weighted signal-to-noise ratio improvement [38] is used, defined as

$$\Delta\text{SNR}_{\text{intellig}} = \sum_i I_i(\text{SNR}_{i,\text{out}} - \text{SNR}_{i,\text{in}}), \tag{59}$$

where the band importance function I_i expresses the importance of the i-th one-third octave band with center frequency f_i^c for intelligibility, $\text{SNR}_{i,\text{out}}$ is the output SNR (in dB) and $\text{SNR}_{i,\text{in}}$ is the input SNR (in dB) in the i-th one third octave band. The center frequencies f_i^c and the values I_i are defined in [39]. The intelligibility weighted signal-to-noise ratio reflects how much intelligibility is improved by the noise reduction algorithms, but does not take into account speech distortion.

To measure the amount of speech distortion, we define the following intelligibility weighted spectral distortion measure

$$\text{SD}_{\text{intellig}} = \sum_i I_i \text{SD}_i \tag{60}$$

with SD_i the average spectral distortion (dB) in i-th one-third band, measured as

$$SD_i = \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} |10 \log_{10} G^s(f)| \, df \Big/ \left[\left(2^{1/6} - 2^{-1/6} \right) f_i^c \right], \tag{61}$$

with $G^s(f)$ the power transfer function of speech from the input to the output of the noise reduction algorithm.

To exclude the effect of the spatial pre-processor, the performance measures are calculated w.r.t. the output of the fixed beamformer.

3.4.3 Experimental results

The impact of the different parameter settings for μ and $\mathbf{w}_0$ on the performance of the SP-SDW-MWF is illustrated for a five noise source scenario. The five noise sources are positioned at angles 75°, 120°, 180°, 240°, 285° w.r.t. the desired source at 0°. To assess the sensitivity of the algorithm against errors in the assumed signal model, the influence of microphone mismatch, e.g., gain mismatch of the second microphone, on the performance is depicted. Among the different possible signal model errors, microphone mismatch was found to be especially harmful to the performance of the GSC in a hearing aid application[17]. In hearing aids, microphones are rarely matched in gain and phase. In [3], gain and phase differences between microphone characteristics of up to 6 dB and 10°, respectively, have been reported.

SP-SDW-MWF without $\mathbf{w}_0$ (SDR-GSC)

Figure 5 plots the improvement $\Delta SNR_{intellig}$ and the speech distortion $SD_{intellig}$ as a function of $\frac{1}{\mu}$ obtained by the SDR-GSC (i.e., the SP-SDW-MWF without filter $\mathbf{w}_0$) for different gain mismatches Υ_2 at the second microphone. *In the absence of microphone mismatch*, the amount of speech leakage into the noise references is limited. Hence, the amount of speech distortion is low for all μ. Since there is still a small amount of speech leakage due to reverberation, the amount of noise reduction and speech distortion slightly decreases for increasing $\frac{1}{\mu}$, especially for $\frac{1}{\mu} > 1$. *In the presence of microphone mismatch*, the amount of speech leakage into the noise references grows. For $\frac{1}{\mu} = 0$ (GSC), the speech gets significantly distorted. Due to the cancellation of the desired signal, also the improvement $\Delta SNR_{intellig}$ degrades. Setting $\frac{1}{\mu} > 0$, improves the performance of the GSC in the presence of model errors without compromising performance in the absence of signal model errors.

SP-SDW-MWF with filter $\mathbf{w}_0$

Figure 6 plots the performance measures $\Delta SNR_{intellig}$ and $SD_{intellig}$ of the *SP-SDW-MWF with filter* $\mathbf{w}_0$. In general, the amount of speech distortion and noise reduction grows for decreasing $\frac{1}{\mu}$. For $\mu = \infty$, all attention is paid to noise reduction. As also illustrated by Figure 6, this results in a total cancellation of the speech and the noise signal and hence degraded performance. *In the absence of model errors*, the

settings $L_0 = 0$ and $L_0 \neq 0$ result - except for $\frac{1}{\mu} = 0$ - in the same $\Delta SNR_{intellig}$ [7], while the distortion for the SP-SDW-MWF with w_0 is higher due to the additional single-channel SDW-MWF. For $L_0 \neq 0$, the performance does -in contrast to $L_0 = 0$ - not degrade due to the microphone mismatch.

Comparison with QIC

Figure 7 depicts the improvement $\Delta SNR_{intellig}$ and the speech distortion $SD_{intellig}$, respectively, of the QIC-GSC as a function of β^2. Like the SDR-GSC, the QIC increases the robustness of the GSC. The QIC is independent of the amount of speech leakage. As a consequence, distortion grows fast with increasing gain deviation. The constraint value β should be chosen so that the maximum permissible speech distortion level is not exceeded for the largest possible model errors. This goes at the expense of reduced noise reduction for small model errors. The SDR-GSC on the other hand, keeps the speech distortion limited for all model errors (see Figure 5). Attention towards speech distortion is increased if the amount of speech leakage grows. As a result, a better noise reduction performance is obtained for small model errors, while guaranteeing sufficient robustness for large model errors. In addition, Figure 6 demonstrates that an additional filter w_0 significantly improves the performance of the SP-SDW-MWF in the presence of signal model errors.

3.5 Conclusion

In the present invention, we established a generalized noise reduction scheme, referred to as *Spatially pre-processed, Speech Distortion Weighted Multi-channel Wiener filter (SP-SDW-MWF)*, that consists of a fixed, spatial pre-processor and an adaptive stage that is based on a SDW-MWF. The new scheme encompasses the GSC and MWF as special cases. In addition, it allows for an in-between solution that can be interpreted as a Speech Distortion Regularized GSC. Depending on the setting of a trade-off *parameter* μ and *the presence or absence of the filter* w_0 *on the speech reference*, the GSC, the SDR-GSC or a (SDW-)MWF is obtained.

In Section 3.2 and Section 3.3, the different parameter settings of the SP-SDW-MWF have been interpreted.

- Without w_0, the SP-SDW-MWF corresponds to a SDR-GSC: the ANC design criterion is supplemented with a regularization term that limits the speech distortion due to signal model errors. The larger $\frac{1}{\mu}$, the smaller the amount of distortion. For $\frac{1}{\mu} = 0$, distortion is ignored completely, which corresponds to the GSC-solution. The SDR-GSC is then an alternative technique to the QIC-GSC to decrease the sensitivity of the GSC to signal model errors. In contrast to the QIC-GSC, the SDR-GSC shifts emphasis towards speech distortion when the amount of speech leakage grows. In the absence of signal model errors, the performance of the GSC is preserved. As a result, a better noise reduction performance is obtained for small model errors, while guaranteeing robustness against large model errors.

[7] For $L_0 \neq 0$, the SNR improvement was larger thanks to the single channel SDW MWF postfilter (see Section 3.3). For other noise sources, e.g., a narrow band noise source, also a better improvement in $SNR_{intellig}$ can be achieved by $L_0 \neq 0$ thanks to the single channel spectral filtering.

- Since the SP-SDW-MWF takes speech distortion explicitly into account, a filter w_0 on the speech reference can be added. It is shown that -in the absence of speech leakage and for infinitely long filter lengths- the SP-SDW-MWF corresponds to a cascade of a SDR-GSC with a SDW-SWF postfilter. In the presence of speech leakage, the SP-SDW-MWF with w_0 tries to preserve its performance: compared to a SDR-GSC with SDW-SWF postfilter, the SP-SDW-MWF then contains extra filtering operations that compensate for the performance degradation of the SDR-GSC with SDW-SWF due to speech leakage. In contrast to the SDR-GSC (and thus also the GSC), performance does not degrade due to microphone mismatch.

In Section 3.4, experimental results for a hearing aid application confirmed the theoretical results of Section 3.2 and Section 3.3. The SP-SDW-MWF indeed increases the robustness of the GSC against signal model errors. Comparison with the widely studied QIC-GSC demonstrated that the SP-SDW-MWF achieves a better noise reduction performance for a given maximum allowable speech distortion level.

4 Third embodiment: Stochastic gradient implementations

In [22, 27] recursive implementations of the MWF have been proposed based on a GSVD or QR decomposition. A subband implementation [28] results in improved intelligibility at a significantly lower cost compared to the fullband approach. These techniques can be extended to implement the SP-SDW-MWF. However, in contrast to the GSC and the QIC-GSC [14], no cheap stochastic gradient based implementation of the SP-SDW-MWF is available. In [25], an LMS based algorithm for the MWF has been developed. The algorithm needs recordings of calibration signals. Since room acoustics, microphone characteristics and the location of the desired speaker change over time, frequent re-calibration is required, making this approach cumbersome and expensive. In [26], an LMS based SDW-MWF has been proposed that avoids the need for calibration signals. The algorithm however relies on some independence assumptions that are not necessarily satisfied. In the present invention, we propose time-domain and frequency-domain stochastic gradient implementations of the SP-SDW-MWF that preserve the benefit of matrix-based SP-SDW-MWF over QIC-GSC. The LMS based SDW-MWF of [26] is modified so that it applies to the SP-SDW-MWF scheme. In addition, other stochastic gradient algorithms are developed that achieve a better performance. Experimental results demonstrate that the proposed stochastic gradient implementation of the SP-SDW-MWF outperforms the SPA, while its computational cost is limited.

This section is organized as follows. Starting from the cost function of the SP-SDW-MWF, a time-domain stochastic gradient algorithm is derived in Section 4.1. Applying the independence assumptions made in [26] results in an LMS based SP-SDW-MWF similar to [26]. To increase convergence and reduce complexity, the stochastic gradient and LMS based algorithm are implemented in the frequency-domain. Both, the stochastic gradient and LMS based algorithm suffer from a large excess error, when applied in highly time-varying noise scenarios. In Section 4.2, we show that the performance of the stochastic gradient algorithm is improved by applying a low pass filter to the part of the gradient estimate that limits speech

distortion. The low pass filtering avoids a highly time-varying distortion of the desired speech component while not degrading the tracking performance needed in time-varying noise scenarios. Section 4.3 compares the performance of the different frequency-domain stochastic gradient algorithms. Experimental results show that the proposed stochastic gradient algorithm preserves the benefit of the SP-SDW-MWF over the QIC-GSC.

4.1 Stochastic gradient algorithm

4.1.1 Derivation

A stochastic gradient algorithm approximates the steepest descent algorithm, using an instantaneous gradient estimate. Given the cost function (41), the steepest descent algorithm iterates as follows[8]

$$\begin{aligned}\mathbf{w}[n+1] &= \mathbf{w}[n] + \frac{\rho}{2}\left(-\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}\right)_{\mathbf{w}=\mathbf{w}[n]} \\ &= \mathbf{w}[n] + \rho\left(\mathcal{E}\{\mathbf{y}^n y_0^{n,*}[k-\Delta]\} - \mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}[k]\}\mathbf{w}[n] - \frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}[k]\}\mathbf{w}[n]\right), \quad (62)\end{aligned}$$

with $\mathbf{w}[k]$, $\mathbf{y}[k] \in \mathbb{C}^{NL\times 1}$, where N denotes the number of input channels to the adaptive filter and L the number of filter taps per channel. Replacing the iteration index n by a time index k and leaving out the expectation values $\mathcal{E}\{.\}$, we obtain the following update equation

$$\boxed{\mathbf{w}[k+1] = \mathbf{w}[k] + \rho\left\{\mathbf{y}^n[k](y_0^{n,*}[k-\Delta] - \mathbf{y}^{n,H}[k]\mathbf{w}[k]) - \underbrace{\frac{1}{\mu}\mathbf{y}^s\mathbf{y}^{s,H}[k]\mathbf{w}[k]}_{\mathbf{r}[k]}\right\}.} \quad (63)$$

For $\frac{1}{\mu} = 0$ and no filtering $\mathbf{w}_0$ on the speech reference, equation (63) reduces to the update formula used in GSC during periods of noise only (i.e., when $y_i[k] = y_i^n[k]$, $i = 0, ..., M-1$). The additional term $\mathbf{r}[k]$ in the gradient estimate limits the speech distortion due to possible signal model errors.

Equation (63) requires knowledge of the correlation matrix $\mathbf{y}^s\mathbf{y}^{s,H}[k]$ or $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}[k]\}$ of the clean speech. In practice, this information is not available. To avoid the need for calibration, *speech + noise* signal vectors $\mathbf{y}_{buf_1}$ are stored into a circular buffer $\mathbf{B}_1 \in \mathbb{R}^{N\times L_{buf_1}}$ during processing as in [26]. During periods of noise only (i.e., when $y_i[k] = y_i^n[k]$, $i = 0, ..., M-1$), the filter $\mathbf{w}$ is updated using the following approximation of the term $\mathbf{r}[k] = \frac{1}{\mu}\mathbf{y}^s\mathbf{y}^{s,H}[k]\mathbf{w}[k]$ in (63)

$$\frac{1}{\mu}\mathbf{y}^s\mathbf{y}^{s,H}[k]\mathbf{w}[k] \approx \frac{1}{\mu}\left(\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H[k] - \mathbf{y}\mathbf{y}^H[k]\right)\mathbf{w}[k], \quad (64)$$

[8] In the sequel the subscripts $0 : M-1$ in the adaptive filter $\mathbf{w}_{0:M-1}$ and the input vector $\mathbf{y}_{0:M-1}$ are omitted for the sake of conciseness.

This results in the update formula

$$\boxed{\mathbf{w}[k+1] = \mathbf{w}[k] + \rho \left\{ \mathbf{y}[k](y_0^*[k-\Delta] - \mathbf{y}^H[k]\mathbf{w}[k]) - \underbrace{\frac{1}{\mu}\left(\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H[k] - \mathbf{y}\mathbf{y}^H[k]\right)\mathbf{w}[k]}_{\mathbf{r}[k]} \right\}} \quad (65)$$

during periods of noise only. In the sequel, a normalized step size ρ is used, i.e.,

$$\rho = \frac{\rho'}{\frac{1}{\mu}\left|\mathbf{y}_{buf_1}^H\mathbf{y}_{buf_1}[k] - \mathbf{y}^H\mathbf{y}[k]\right| + \mathbf{y}^H\mathbf{y}[k] + \delta}, \quad (66)$$

where δ is a very small constant. The absolute value $\left|\mathbf{y}_{buf_1}^H\mathbf{y}_{buf_1} - \mathbf{y}^H\mathbf{y}\right|$ has been inserted to guarantee a positive valued estimate of the clean speech energy $\mathbf{y}^{s,H}\mathbf{y}^s[k]$. Additional storage of noise only vectors $\mathbf{y}_{buf_2} \in \mathbb{C}^{ML\times 1}$ in a second buffer $\mathbf{B}_2 \in \mathbb{R}^{M\times L_{buf_2}}$ allows to adapt $\mathbf{w}$ also during periods of speech + noise, using

$$\mathbf{w}[k+1] = \mathbf{w}[k] + \rho\left\{\mathbf{y}_{buf_2}(y_{0,buf_2}^*[k-\Delta] - \mathbf{y}_{buf_2}^H\mathbf{w}[k]) + \frac{1}{\mu}\left(\mathbf{y}_{buf_2}\mathbf{y}_{buf_2}^H[k] - \mathbf{y}\mathbf{y}^H[k]\right)\mathbf{w}[k]\right\} \quad (67)$$

with

$$\rho = \frac{\rho'}{\frac{1}{\mu}\left|\mathbf{y}^H\mathbf{y} - \mathbf{y}_{buf_2}^H\mathbf{y}_{buf_2}\right| + \mathbf{y}_{buf_2}^H\mathbf{y}_{buf_2} + \delta}. \quad (68)$$

In the sequel, we will - for reasons of conciseness- only consider the update procedure of the time-domain stochastic gradient algorithms during noise only, hence, $\mathbf{y}[k] = \mathbf{y}^n[k]$. The extension towards updating during speech + noise periods with the use of a second, noise only buffer $\mathbf{B}_2$ is straightforward: the equations are found by replacing the noise-only, input vectors $\mathbf{y}[k]$ by $\mathbf{y}_{buf_2}[k]$ and the speech + noise vectors $\mathbf{y}_{buf_1}[k]$ by the input speech + noise vector $\mathbf{y}[k]$.

Using[9]

$$\mathbf{w}_{opt} = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H\} + (1-\frac{1}{\mu})\mathcal{E}\{\mathbf{y}\mathbf{y}^H\}\right)^{-1}\mathcal{E}\{\mathbf{y}^H y_0^*[k-\Delta]\}, \quad (69)$$

where $\mathbf{y}$ is a noise-only vector, and (65) it can be shown that

$$\mathcal{E}\{\mathbf{w}[k+1] - \mathbf{w}_{opt}\} = \left(\mathbf{I} - \rho\mathcal{E}\{\frac{1}{\mu}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H + (1-\frac{1}{\mu})\mathbf{y}\mathbf{y}^H\}\right)^{k+1}\mathcal{E}\{\mathbf{w}[0] - \mathbf{w}_{opt}\}. \quad (70)$$

Hence, the algorithm (65)-(67) is convergent in the mean provided that the step size ρ is smaller than $\frac{2}{\lambda_{\max}}$ with $\lambda_{\max}$ the maximum eigenvalue of $\mathcal{E}\{\frac{1}{\mu}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H + (1-\frac{1}{\mu})\mathbf{y}\mathbf{y}^H\}$. The similarity of (65) with standard NLMS let us presume that setting $\rho < \frac{2}{\sum_i^{NL}\lambda_i}$, with λ_i, $i = 1, \ldots, NL$ the eigenvalues of $\mathcal{E}\{\frac{1}{\mu}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H +$

[9]When the second order statistics of the noise are short-term stationary, $\mathbf{w}_{opt}$ equals to (36).

$(1-\frac{1}{\mu})\mathbf{y}\mathbf{y}^H\} \in \mathbb{R}^{NL\times NL}$, or -in case of FIR filters- setting

$$\rho < \frac{2}{\frac{1}{\mu}L\sum_{i=M-N}^{M-1}\mathcal{E}\{y_{i,buf_1}^2[k]\} + (1-\frac{1}{\mu})L\sum_{i=M-N}^{M-1}\mathcal{E}\{y_i^2[k]\}} \tag{71}$$

guarantees convergence in the mean square. Equation (71) explains the normalization (66) and (68) for the step size ρ.

However, since generally

$$\mathbf{y}\mathbf{y}^H[k] \neq \mathbf{y}_{buf_1}^n\mathbf{y}_{buf_1}^{n,H}[k], \tag{72}$$

the instantaneous gradient estimate in (65) is -compared to (63)- additionally perturbed by

$$\frac{1}{\mu}\left(\mathbf{y}\mathbf{y}^H[k] - \mathbf{y}_{buf_1}^n\mathbf{y}_{buf_1}^{n,H}[k]\right)\mathbf{w}[k], \tag{73}$$

for $\mu \neq \infty$. Hence, for $\mu \neq \infty$, the update equation (65)-(67) suffers from a larger residual excess error than (63). The additional excess error grows for decreasing μ, increasing step size ρ and increasing vector length $L.N$ of the vector $\mathbf{y}$ with L the filter length per channel and N the number of inputs to the adaptive filter. It is expected to be especially large for highly time-varying noise, e.g., multi-talker babble noise.

4.1.2 NLMS based algorithm

In [26], an LMS based implementation of the SDW-MWF has been proposed. Besides (64), some additional, independence assumptions are made. Applying these assumptions to (65)-(67), results in an LMS based implementation of the SP-SDW-MWF similar to [26]. Assuming that

$$\sqrt{\frac{1}{\mu-1}}\mathbf{y}_{buf_1}[k]y_0^*[k-\Delta] = 0 \tag{74}$$

$$\sqrt{\frac{1}{\mu}\left(1-\frac{1}{\mu}\right)}\left(\mathbf{y}[k]\mathbf{y}_{buf_1}^H[k] + \mathbf{y}_{buf_1}[k]\mathbf{y}^H[k]\right) = 0, \tag{75}$$

hold, with k and l different time instants, (65) can be simplified to

$$\boxed{\mathbf{w}[k+1] = \mathbf{w}[k] + \frac{\rho'}{\mathbf{x}^H[k]\mathbf{x}[k]+\delta}\mathbf{x}[k](d^*[k] - \mathbf{x}^H[k]\mathbf{w}[k])} \tag{76}$$

where

$$d[k] = y_0[k-\Delta]\frac{1}{\sqrt{1-\frac{1}{\mu}}}; \quad \mathbf{x}[k] = \sqrt{1-\tfrac{1}{\mu}}\mathbf{y}[k] + \sqrt{\tfrac{1}{\mu}}\mathbf{y}_{buf_1}[k] \tag{77}$$

during periods of noise only (i.e., $\mathbf{y}[k] = \mathbf{y}^n[k]$). During speech + noise (i.e., $\mathbf{y}[k] = \mathbf{y}^s[k] + \mathbf{y}^n[k]$), $d[k]$ and $\mathbf{x}[k]$ in (76) are set to

$$d[k] = y_{0,buf_2}[k-\Delta]\frac{1}{\sqrt{1-\frac{1}{\mu}}}; \mathbf{x}[k] = \sqrt{1-\frac{1}{\mu}}\mathbf{y}_{buf_2}[k] + \sqrt{\frac{1}{\mu}}\mathbf{y}[k]. \tag{78}$$

Equations (74) and (75) assume that - besides speech and noise vectors - also noise vectors at different time instants are mutually uncorrelated. In practice, (74) and (75) do not hold, especially for large $\sqrt{\frac{1}{\mu-1}}$ and $\sqrt{\frac{1}{\mu}\left(1-\frac{1}{\mu}\right)}$, i.e. for $\mu \rightarrow 1$. Hence, compared to (65)-(67), performance is expected to be worse. In addition, equations (76)-(78) can - in contrast to (65)- not be applied for $\mu \leq 1$. Compared to (65) no significant complexity reduction is achieved. The LMS based updating (76) requires $4NL + 3$ Multiply-Accumulate (MAC) per sample[10], whereas update formula (65) requires $(4NL + 5)$ MAC per sample. The computation of the normalized step size in (76) requires $NL + 2$ less MAC per sample than in (65).

4.1.3 Frequency-domain implementation

As stated before, the stochastic gradient algorithms (65)-(67) and (76) are expected to suffer from a large excess error for large $\frac{\rho'}{\mu}$ and/or highly time-varying noise, due to a large difference between the rank-one noise correlation matrices $\mathbf{y}^n\mathbf{y}^{n,H}[k]$ measured at different time instants k. The gradient estimate can be improved by replacing

$$\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H[k] - \mathbf{y}\mathbf{y}^H[k] \tag{79}$$

in (65) with the time-average

$$\frac{1}{K}\sum_{l=k-K+1}^{k}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H[l] - \frac{1}{K}\sum_{l=k-K+1}^{k}\mathbf{y}\mathbf{y}^H[l], \tag{80}$$

where $\frac{1}{K}\sum_{l=k-K+1}^{k}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H[l]$ is updated during periods of speech + noise and $\frac{1}{K}\sum_{l=k-K+1}^{k}\mathbf{y}\mathbf{y}^H[l]$ during periods of noise only. However, this would require expensive matrix operations. A block-based implementation intrinsically performs this averaging:

$$\begin{aligned}\mathbf{w}[(k+1)K] &= \mathbf{w}[kK] + \frac{\rho}{K}\left[\sum_{i=0}^{K-1}\mathbf{y}[kK+i]\left(y_0^*[kK+i-\Delta] - \mathbf{y}^H[kK+i]\mathbf{w}[kK]\right)\right. \\ &\left. -\frac{1}{\mu}\sum_{i=0}^{K-1}\left(\mathbf{y}_{buf_1}[kK+i]\mathbf{y}_{buf_1}^H[kK+i] - \mathbf{y}[kK+i]\mathbf{y}^H[kK+i]\right)\mathbf{w}[kK]\right].\end{aligned} \tag{81}$$

[10]Note that the output $y_0[k-\Delta] - \mathbf{w}^H\mathbf{y}[k]$ of the algorithm still has to be computed.

The gradient and hence also $\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H[k] - \mathbf{y}\mathbf{y}^H[k]$ is averaged over K iterations prior to make adjustments to w. This goes at the expense of a reduced (i.e. by a factor K) convergence rate.

The block-based implementation is computationally more efficient when it is implemented in the frequency-domain, especially for large filter lengths. In addition, in a frequency-domain implementation, each frequency bin gets its own step size, resulting in faster convergence compared to a time-domain implementation while not degrading the time-domain MSE. Although the frequency and time-domain implementation obtain the same MSE, the improvement in $\text{SNR}_{\text{intellig}}$, which is determined by the excess errors in each frequency bin, may be different. In a *time-domain implementation*, one common step size ρ is used for the different frequency bins. The convergence rate depends on the eigenvalue spread of the correlation matrix of the input signals to the adaptive filter and hence on the power spectrum of the input signal. In frequency bins with little power this common step size will be smaller than in the frequency-domain approach, resulting in slower convergence and less excess error in that bin. In frequency bins with large power on the other hand, this common step size will be larger than in the frequency-domain approach, resulting in larger LMS excess error in that frequency bin. Hence, in a time-domain implementation, the power spectrum of the input signals not only determines the convergence rate but also the improvement $\Delta\text{SNR}_{\text{intellig}}$. In a *frequency-domain implementation*, the step size is normalized in each frequency bin, so that the different bins have a similar convergence rate and hence also excess error. Hence, the SNR improvement in each frequency bin is more controlled (i.e. less dependent on the input power spectrum). Since signal model errors (e.g., microphone mismatch) modify the power spectrum of the noise references and hence, the convergence rate and improvement $\Delta\text{SNR}_{\text{intellig}}$ of a time-domain implementation, frequency-domain implementations are more appropriate to evaluate the performance of the algorithms for different signal model errors.

Algorithm 1 and Algorithm 2 summarize a frequency-domain implementation based on overlap-save of (65)-(67) and (76), respectively. Algorithm 1 requires $(3N + 4)$ FFTs of length $2L$ and algorithm 2 $(3N + 3)$ FFTs. By storing the FFT-transformed speech + noise and noise-only vectors in the buffers[11] $\mathbf{B}_1 \in \mathbb{C}^{N \times L_{buf_1}}$ and $\mathbf{B}_2 \in \mathbb{C}^{N \times L_{buf_2}}$, respectively, instead of storing the time-domain vectors, N FFT operations have been saved. When adapting during speech + noise, also the time-domain vector

$$\left[\, y_0[kL - \Delta] \quad \cdots \quad y_0[kL - \Delta + L - 1] \,\right]^T \tag{82}$$

should then be stored in an additional buffer $\mathbf{B}_{2,0} \in \mathbb{R}^{1 \times \frac{L_{buf_2}}{2}}$ during periods of noise-only, which -for $N = M$- results in an additional storage of $\frac{L_{buf_2}}{2}$ words compared to when the time-domain vectors are stored into the buffers $\mathbf{B}_1$ and $\mathbf{B}_2$.

Remark : In algorithm 1 and 2 a common trade-off parameter μ is used in all frequency bins. Alternatively, a different setting for μ can be used in different frequency bins. E.g. for SP-SDW-MWF with $\mathbf{w}_0 = \mathbf{0}$, μ could be set to ∞ at those frequencies where the GSC is sufficiently robust, e.g., for small-sized arrays at

[11] Since the input signals are real, half of the FFT components are complex-conjugated. Hence, in practice only half of the complex FFT components have to be stored in memory.

Algorithm 1 Frequency domain stochastic gradient SP-SDW-MWF based on overlap-save.

Initialization:

$\mathbf{W}_i[0] = [\ 0 \quad \cdots \quad 0\ ]^T$; $i = M-N, ..., M-1$
$P_m[0] = \delta_m$; $m = 0, ..., 2L-1$

Matrix definitions:

$$\mathbf{g} = \begin{bmatrix} \mathbf{I}_L & \mathbf{0}_L \\ \mathbf{0}_L & \mathbf{0}_L \end{bmatrix}; \mathbf{k} = [\ \mathbf{0} \quad \mathbf{I}_L\ ]; \mathbf{F} = 2L \times 2L \text{ DFT matrix}$$

For each new block of NL input samples:

- *If noise detected:*

  1. $\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T$, $i = M-N, ..., M-1 \rightarrow$ noise buffer $\mathbf{B}_2$
     $[\ y_0[kL-\Delta] \quad ... \quad y_0[kL-\Delta+L-1]\ ]^T \rightarrow$ noise buffer $\mathbf{B}_{2,0}$
  2. $\mathbf{Y}_i^n[k] = \text{diag}\left\{\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T\right\}$, $i = M-N, ..., M-1$
     $\mathbf{Y}_i[k] = \text{diag}\left\{[\ \mathbf{B}_1(i,0) \quad ... \quad \mathbf{B}_1(i,2L-1)\ ]^T\right\}$, $i = M-N, ..., M-1$
     cyclically shift each row i of $\mathbf{B}_1$ over $2L$ samples, $i = M-N, ..., M-1$
     $\mathbf{d}[k] = [\ 0 \quad \cdots \quad 0 \quad y_0[kL-\Delta] \quad \cdots \quad y_0[kL-\Delta+L-1\ ]]^T$

- *If speech detected:*

  1. $\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T$, $i = M-N, ..., M-1 \rightarrow$ speech+noise buffer $\mathbf{B}_1$
  2. $\mathbf{Y}_i^n[k] = \text{diag}\left\{[\ \mathbf{B}_2(i,0) \quad ... \quad \mathbf{B}_2(i,2L-1)\ ]^T\right\}$, $i = M-N, ..., M-1$
     cyclically shift each row i of $\mathbf{B}_2$ over $2L$ samples, $i = M-N, ..., M-1$
     $\mathbf{Y}_i[k] = \text{diag}\left\{\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T\right\}$, $i = M-N, ..., M-1$
     $\mathbf{d}[k] = [\ 0 \quad \cdots \quad 0 \quad \mathbf{B}_{2,0}(1,0) \quad \cdots \quad \mathbf{B}_{2,0}(1,L-1)\ ]^T$
     cyclically shift $\mathbf{B}_{2,0}$ over L samples

- *Update formula:*

  1. $\mathbf{e}_1[k] = \mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j^n[k]\mathbf{W}_j^*[k] = \mathbf{y}_{\text{out},1}$
     $\mathbf{e}[k] = \mathbf{d}[k] - \mathbf{e}_1[k]$
     $\mathbf{e}_2[k] = \mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j[k]\mathbf{W}_j^*[k] = \mathbf{y}_{\text{out},2}$

     $\mathbf{E}_1[k] = \mathbf{F}\mathbf{k}^T\mathbf{e}_1[k]$; $\mathbf{E}_2[k] = \mathbf{F}\mathbf{k}^T\mathbf{e}_2[k]$; $\mathbf{E}[k] = \mathbf{F}\mathbf{k}^T\mathbf{e}[k]$
  2. $\Lambda[k] = \frac{2\rho'}{L}\text{diag}\ \{P_0^{-1}[k], ..., P_{2L-1}^{-1}[k]\}$
     $P_m[k] = \gamma P_m[k-1] + (1-\gamma)\left(\sum_{j=M-N}^{M-1}|\mathbf{Y}_{j,m}^n|^2 + \frac{1}{\mu}\left|\sum_{j=M-N}^{M-1}\left(|\mathbf{Y}_{j,m}|^2 - |\mathbf{Y}_{j,m}^n|^2\right)\right|\right)$
  3. $\mathbf{W}_i[k+1] = \mathbf{W}_i[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\Lambda[k]\left\{\mathbf{Y}_i^n[k]\mathbf{E}^*[k] - \frac{1}{\mu}\left(\mathbf{Y}_i\mathbf{E}_2^*[k] - \mathbf{Y}_i^n\mathbf{E}_1^*[k]\right)\right\}$, $i = M-N, ..., M-1$

- Output: $\mathbf{y}_0[k] = [\ y_0[kL-\Delta] \quad \cdots \quad y_0[kL-\Delta+L-1]\ ]^T$
  - If noise detected: $\mathbf{y}_{\text{out}}[k] = \mathbf{y}_0[k] - \mathbf{y}_{\text{out},1}[k]$
  - If speech detected: $\mathbf{y}_{\text{out}}[k] = \mathbf{y}_0[k] - \mathbf{y}_{\text{out},2}[k]$

Algorithm 2 Frequency domain NLMS based SP-SDW-MWF based on overlap-save.

Initialization:

$\mathbf{W}_i[0] = [\ 0 \quad \cdots \quad 0\ ]^T,\ i = M-N, ..., M-1$
$P_m[0] = \delta_m,\ m = 0, ..., 2L-1$

Matrix definitions:

$\mathbf{g} = \begin{bmatrix} \mathbf{I}_L & \mathbf{0}_L \\ \mathbf{0}_L & \mathbf{0}_L \end{bmatrix}$; $\mathbf{k} = [\ \mathbf{0} \quad \mathbf{I}_L\ ]$; $\mathbf{F} = 2L \times 2L$ DFT matrix

For each new block of NL input samples:

- *If noise detected:*
  1. $\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T,\ i = M-N, ..., M-1 \rightarrow$ noise buffer $\mathbf{B}_2$
     $[\ y_0[kL-\Delta] \quad ... \quad y_0[kL-\Delta+L-1]\ ]^T \rightarrow$ noise buffer $\mathbf{B}_{2,0}$
  2. $\mathbf{Y}_i[k] = \text{diag}\left\{\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T\right\},\ i = M-N, ..., M-1$
     $\mathbf{Y}_{i,buf_1}[k] = \text{diag}\left\{[\ \mathbf{B}_1(i,0) \quad ... \quad \mathbf{B}_1(i,2L-1)\ ]^T\right\},\ i = M-N, ..., M-1$
     $\mathbf{X}_i[k] = \sqrt{1-\frac{1}{\mu}}\mathbf{Y}_i[k] + \sqrt{\frac{1}{\mu}}\mathbf{Y}_{i,buf_1}[k],\ i = M-N, ..., M-1$
     cyclically shift each row i of buffer $\mathbf{B}_1$ over $2L$ samples
     $\mathbf{d}[k] = \frac{1}{\sqrt{1-\frac{1}{\mu}}}[\ 0 \quad \cdots \quad 0 \quad y_0[kL-\Delta] \quad \cdots \quad y_0[kL-\Delta+L-1\ ]]^T$
- *If speech detected:*
  1. $\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T,\ i = M-N, ..., M-1 \rightarrow$ speech+noise buffer $\mathbf{B}_1$
  2. $\mathbf{Y}_i[k] = \text{diag}\left\{\mathbf{F}[\ y_i[kL-L] \quad ... \quad y_i[kL+L-1]\ ]^T\right\},\ i = M-N, ..., M-1$
     $\mathbf{Y}_{i,buf_2}[k] = \text{diag}\left\{[\ \mathbf{B}_2(i,0) \quad ... \quad \mathbf{B}_2(i,2L-1)\ ]^T\right\},\ i = M-N, ..., M-1$
     $\mathbf{X}_i[k] = \sqrt{1-\frac{1}{\mu}}\mathbf{Y}_{i,buf_2}[k] + \sqrt{\frac{1}{\mu}}\mathbf{Y}_i[k],\ i = M-N, ..., M-1$
     cyclically shift each row i of buffer $\mathbf{B}_2$ over $2L$ samples
     $\mathbf{d}[k] = \frac{1}{\sqrt{1-\frac{1}{\mu}}}[\ 0 \quad \cdots \quad 0 \quad \mathbf{B}_{2,0}(1,0) \quad \cdots \quad \mathbf{B}_{2,0}(1,L-1)\ ]^T$
     cyclically shift $\mathbf{B}_{2,0}$ over L samples
- *Update formula:*
  1. $\mathbf{E}[k] = \mathbf{F}\mathbf{k}^T\left(\mathbf{d} - \mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{X}_j[k]\mathbf{W}_j^*[k]\right)$
  2. $\Lambda[k] = \frac{2\rho'}{L}\text{diag}\left\{P_0^{-1}[k], ..., P_{2L-1}^{-1}[k]\right\}$
     $P_m[k] = \gamma P_m[k-1] + (1-\gamma)\left(\sum_{j=M-N}^{M-1}|\mathbf{X}_{j,m}|^2\right)$
  3. $\mathbf{W}_i[k+1] = \mathbf{W}_i[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\Lambda[k]\mathbf{X}_i[k]\mathbf{E}^*[k],\ i = M-N, ..., M-1$
- Output: $\mathbf{y}_0[k] - \mathbf{k}\mathbf{F}^{-1}\sum_{i=M-N}^{M-1}\mathbf{Y}_i[k]\mathbf{W}_i^*[k]$
  $\mathbf{y}_0[k] = [\ y_0[kL-\Delta] \quad \cdots \quad y_0[kL-\Delta+L-1]\ ]^T$

high frequencies. In that case, only a few frequency components of $\mathbf{Y}_i$ should be stored in the speech + noise buffer.

4.2 Improvement of stochastic gradient algorithm

To achieve a reliable estimate (80) of the average correlation matrix $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ in highly time-varying[12] noise scenarios (e.g. multi-talker babble), K should be much larger than LN. Hence, the averaging in the block-based[13] or frequency-domain implementation proposed in Section 4.1, does not suffice to obtain a good estimate for $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$. In this Section, we show that the performance of the stochastic gradient algorithm is improved by applying a low pass filter to the part of the gradient estimate that takes speech distortion into account, i.e., the term $\mathbf{r}[k]$ in (65). The low pass filtering avoids a highly time-varying distortion of the desired speech component while not degrading the tracking performance needed in non-stationary noise scenarios.

4.2.1 Concept

Define $\mathbf{w}_s$ as[14]

$$\mathbf{w}_s = \mathbf{w} - \mathbf{w}_n \tag{83}$$

$$\mathbf{w}_s \in \text{Range}\{\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}\} \tag{84}$$

$$\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}\mathbf{w}_n = \mathbf{0}. \tag{85}$$

Then, the desired speech component $z^s[k]$ at the output equals

$$z^s[k] = y_0^s[k-\Delta] - \mathbf{w}_s^H\mathbf{y}^s[k]. \tag{86}$$

Assume that $\mathbf{w}_s$ varies slowly in time. This is desired since a fast changing $\mathbf{w}_s$ results in a highly time-varying distortion of the desired speech, and may thus harm sound quality. In addition, in hearing aid applications the average correlation matrix $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ is slowly time-varying as microphone characteristics, room acoustics and the average desired speaker position do not change quickly in time. Fast changes in the noise scenario can be tracked by the filter $\mathbf{w}_n$. This will be illustrated in Section 4.3.

Then,

$$\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}\mathbf{w}[k] = \mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}\mathbf{w}_s \tag{87}$$

can be approximated by[15]

$$\mathcal{E}\{\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H - \mathbf{y}\mathbf{y}^H\}\mathbf{w}_s = \mathcal{E}\{\left(\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H - \mathbf{y}\mathbf{y}^H\right)\mathbf{w}_s\}, \tag{88}$$

[12] Like for the QR and GSVD based algorithms, we assume short-term stationarity of the second order statistics of the noise, so that $\mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}\} \approx \mathcal{E}\{\mathbf{y}_{buf_1}^n\mathbf{y}_{buf_1}^{n,H}\}$. The first and higher order statistics are allowed to vary faster in time.

[13] A large $K \gg LN$ in block-LMS would result in a too slow convergence rate.

[14] $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ is rank deficient when the speech leakage $\mathbf{y}_s$ in the noise references does not cover the whole frequency spectrum or when the number of inputs N to the adaptive filter exceeds 1 and the direct-to-reverberant ratio of the desired speech is high.

[15] Just like for the matrix based algorithms, the noise correlation matrix $\mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}\}$ is assumed to be short-term stationary so that it can be estimated during periods of noise only.

where $\mathbf{y}$ is a vector during noise only. Using the independence assumption [40]

$$\mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}[k]\mathbf{w}_n[k]\} = \mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}[k]\}\mathcal{E}\{\mathbf{w}_n[k]\} \tag{89}$$

and $\mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}\} = \mathcal{E}\{\mathbf{y}^n_{buf_1}\mathbf{y}^{n,H}_{buf_1}\}$, we find that

$$\mathcal{E}\{\left(\mathbf{y}_{buf_1}\mathbf{y}^H_{buf_1} - \mathbf{y}\mathbf{y}^H\right)\mathbf{w}_s\} = \mathcal{E}\{\left(\mathbf{y}_{buf_1}\mathbf{y}^H_{buf_1} - \mathbf{y}\mathbf{y}^H\right)\mathbf{w}[k]\}. \tag{90}$$

Replacing the expectation value by time averaging, $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}\mathbf{w}[k]$ can be estimated as

$$\frac{1}{K}\sum_{l=k-K}^{l=k}\left(\mathbf{y}_{buf_1}\mathbf{y}^H_{buf_1}[l] - \mathbf{y}\mathbf{y}^H[l]\right)\mathbf{w}[l] \tag{91}$$

during noise only[16]. The value K determines the convergence rate of the filter $\mathbf{w}_s$.

Remark: In order to obtain a good estimate of $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$, the long-term averaged noise correlation matrices $\frac{1}{K}\sum_{l=k-K}^{l=k}\mathbf{y}\mathbf{y}^H[l]$ and $\frac{1}{K}\sum_{l=k-K}^{l=k}\mathbf{y}^n_{buf_1}\mathbf{y}^{n,H}_{buf_1}[l]$ should not differ too much from each other. This does not requires that the second order statistics of the noise source are stationary for about K time samples. It suffices that they are short-term stationary so that they can be estimated during noise only periods.

The averaging operation (91) is performed by applying the following low pass filter to the term $\mathbf{r}[k] = \frac{1}{\mu}\left(\mathbf{y}_{buf_1}\mathbf{y}^H_{buf_1} - \mathbf{y}\mathbf{y}^H\right)\mathbf{w}[k]$ in (65):

$$\boxed{\mathbf{r}[k] = \tilde{\lambda}\mathbf{r}[k-1] + (1-\tilde{\lambda})\frac{1}{\mu}\left(\mathbf{y}_{buf_1}\mathbf{y}^H_{buf_1} - \mathbf{y}\mathbf{y}^H\right)\mathbf{w}[k],} \tag{92}$$

where $\tilde{\lambda} < 1$. This corresponds to an averaging window K of about $\frac{1}{1-\tilde{\lambda}}$ samples. The normalized step size ρ is modified into

$$\rho = \frac{\rho'}{r_{avg}[k] + \mathbf{y}^H\mathbf{y} + \delta} \tag{93}$$

$$r_{avg}[k] = \tilde{\lambda}r_{avg}[k-1] + (1-\tilde{\lambda})\frac{1}{\mu}\left|\mathbf{y}^H_{buf_1}\mathbf{y}_{buf_1} - \mathbf{y}^H\mathbf{y}\right|. \tag{94}$$

Compared to (65), (92) requires $3NL - 1$ additional MAC and extra storage of a $NL \times 1$ vector $\mathbf{r}$.

4.2.2 Frequency-domain

Equation (92) can be extended to the frequency-domain. The update equation for $\mathbf{W}_i[k+1]$ in algorithm 1 then becomes:

[16] As also mentioned in Section 4.1, the noise-only vector $\mathbf{y}[k]$ should be replaced by $\mathbf{y}_{buf_2}[k]$ and the speech + noise vector $\mathbf{y}_{buf_1}[k]$ by $\mathbf{y}[k]$ when adapting during periods of speech + noise.

$$\begin{aligned} \mathbf{W}_i[k+1] &= \mathbf{W}_i[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\Lambda[k]\left(\mathbf{Y}_i^n[k]\mathbf{E}^*[k] - \mathbf{R}_i[k]\right); \\ \mathbf{R}_i[k] &= \lambda\mathbf{R}_i[k-1] + (1-\lambda)\frac{1}{\mu}\left(\mathbf{Y}_i[k]\mathbf{E}_2^*[k] - \mathbf{Y}_i^n[k]\mathbf{E}_1^*[k]\right) \end{aligned} \tag{95}$$

with

$$\mathbf{E}[k] = \mathbf{F}\mathbf{k}^T\left(\mathbf{y}_0^n - \mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j^n[k]\mathbf{W}_j^*[k]\right); \tag{96}$$

$$\mathbf{E}_1[k] = \mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j^n[k]\mathbf{W}_j^*[k]; \tag{97}$$

$$\mathbf{E}_2[k] = \mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j[k]\mathbf{W}_j^*[k]. \tag{98}$$

and $\rho[k]$ computed as follows:

$$\begin{aligned} \rho[k] &= \frac{2\rho'}{L}\text{diag}\ \{P_0^{-1}[k], ..., P_{2L-1}^{-1}[k]\} \\ P_m[k] &= P_{1,m}[k] + P_{2,m}[k] \\ P_{1,m}[k] &= \gamma P_{1,m}[k-1] + (1-\gamma)\sum_{j=M-N}^{M-1}\left|\mathbf{Y}_{j,m}^n\right|^2 \\ P_{2,m}[k] &= \lambda P_{2,m}[k-1] + (1-\lambda)\frac{1}{\mu}\left|\sum_{j=M-N}^{M-1}\left(|\mathbf{Y}_{j,m}|^2 - \left|\mathbf{Y}_{j,m}^n\right|^2\right)\right|. \end{aligned}$$

Compared to algorithm 1, (95)-(98) requires one extra $2L$−point FFT and $8NL - 2N - 2L$ extra MAC per L samples and additional memory storage of a $2NL \times 1$ real data vector. To obtain the same time constant in the averaging operation as in the time-domain version with $K = 1$, λ should equal $\bar{\lambda}^L$.

Experimental results in Section 4.3 will show that the performance of the stochastic gradient algorithm significantly improves by the low pass filter, especially for large λ.

4.2.3 Complexity of different stochastic gradient algorithms

Table 1 summarizes the computational complexity (expressed as the number of real multiply-accumulate[17] (MAC), divisions (D), square roots (Sq) and absolute values (Abs)) of the time-domain (TD) and frequency-domain (FD) Stochastic Gradient (SG) and NLMS based algorithms. Comparison is made with standard NLMS and the NLMS based SPA. We assume that one complex multiplication is equivalent to 4 real multiplications and 2 real additions. A $2L$-point FFT of a real input vector requires $2L\log_2 2L$ real MAC

[17] counted as the number of multiply-accumulate, additions and multiplications.

(assuming radix-2 FFT algorithms).

Table 1 indicates that the TD-SG without filter w_0 and the SPA are about twice as complex as the standard ANC. When applying a Low Pass filter (LP) to the regularization term, the TD-SG algorithm has about three times the complexity of the ANC. The increase in complexity of the frequency-domain implementations is less.

Table 1: Computational complexity of TD and FD-NLMS and stochastic gradient algorithms (expressed as number of real MAC, divisions (D), absolute values (Abs) and square roots (Sq) per sample)

Algorithm		update formula	adaptation of step size
TD	NLMS ANC	$(2M-2)L+1)$ MAC	1 D+$(M-1)L$ MAC
	NLMS based SPA	$(4(M-1)L+1)$ MAC+1 D+1 Sq	1 D+$(M-1)L$ MAC
	SG	$(4NL+5)$ MAC	1 D + 1 Abs+$(2NL+2)$ MAC
	NLMS based algorithm	$(4NL+3)$ MAC	1 D+NL MAC
	SG with LP	$(7NL+4)$ MAC	1 D+1 abs+$(2NL+4)$ MAC
FD	NLMS ANC	$(10M-7-\frac{4(M-1)}{L})+(6M-2)\log_2 2L$ MAC	1 D+$(2M+2)$ MAC
	NLMS based SPA	$(14M-11-\frac{4(M-1)}{L}+(6M-2)\log_2 2L$ MAC +1/L Sq+1/L D	1 D+$(2M+2)$ MAC
	SG	$(18N+6-\frac{8N}{L})+(6N+8)\log_2 2L$ MAC	1 D+1 abs + $(4N+4)$ MAC
	NLMS based algorithm	$(16N+4-\frac{6N}{L})+(6N+6)\log_2 2L$ MAC	1 D+$(2N+2)$ MAC
	SG with LP	$(26N+4-\frac{10N}{L})+(6N+10)\log_2 2L$ MAC	1 D+1 Abs+$(4N+6)$ MAC

Remark: In Table 1 and Figure 8, the complexity of time-domain and frequency-domain NLMS ANC and NLMS based SPA represents the complexity when the adaptive filter is only updated during noise only. If the adaptive filter is also updated during speech + noise using data from a noise buffer, the time-domain implementations require NL additional MAC per sample and the frequency-domain implementations require 2 additional FFT and $(4L(M-1)-2(M-1)+L)$ MAC per L samples.

As an illustration, Figure 8 plots the complexity (expressed as the number of Mega operations per second (Mops)) of the time-domain and frequency-domain stochastic gradient algorithm with LP filtering as a function of L for $M = 3$ and a sampling frequency $f_s = 16$ kHz. Comparison is made with the NLMS-based ANC of the GSC and the SPA. The complexity of the FD SPA is not depicted, since for small M, it is comparable to the cost of the FD-NLMS ANC. For $L > 8$, the frequency-domain implementations result in a significantly lower complexity compared to their time-domain equivalents. The computational cost of the FD stochastic gradient algorithm with LP is limited, making it a good alternative to the SPA for implementation in hearing aids.

4.3 Experimental results

In this Section, we evaluate the performance of the different FD stochastic gradient algorithms based on experimental results for a hearing aid application. Comparison is made with the FD-NLMS based SPA. For

a fair comparison, the FD-NLMS based SPA is -like the stochastic gradient algorithms- also adapted during speech + noise using data from a noise buffer.

4.3.1 Set-up

A three-microphone Behind-The-Ear (BTE) hearing aid with three omnidirectional microphones (Knowles FG-3452) has been mounted on a dummy head in an office room. The interspacing d between the first and the second microphone is about $d = 1$ cm and the interspacing between the second and third microphone about 1.5 cm. The reverberation time $T_{60\,\mathrm{dB}}$ is about 700 ms for a speech weighted noise. The desired speech signal and the noise signals are uncorrelated. The desired speech source consists of sentences spoken by a male speaker. Both the speech and the noise signal have a level of 70 dB SPL at the center of the head. The desired speech source and noise sources are positioned at a distance of 1 meter from the head: the speech source in front of the head, the noise sources at an angle θ w.r.t. the speech source. For evaluation purposes, the speech and noise signal have been recorded separately.

The microphone signals are pre-whitened prior to processing to improve intelligibility [37], and the output is accordingly de-whitened. In the experiments, the microphones have been calibrated by means of recordings of an anechoic speech weighted noise signal positioned at 0° measured while the microphone array was mounted on the head. A delay-and-sum beamformer is used as a fixed beamformer, since -in case of small microphone interspacing - it is robust to model errors. The blocking matrix $\mathbf{B}$ pairwise subtracts the time aligned calibrated microphone signals.

The performance of the FD stochastic gradient algorithms is evaluated for a filter length $L = 32$ taps per channel, $\rho' = 0.8$ and $\gamma = 0$. To exclude the effect of the spatial pre-processor, the performance measures are calculated w.r.t. the output of the fixed beamformer. The sensitivity of the algorithms against errors in the assumed signal model is illustrated for microphone mismatch, e.g., a gain mismatch $\Upsilon_2 = 4$ dB of the second microphone. Among the different possible signal model errors, especially microphone mismatch was found to be harmful to the performance of the GSC in a hearing aid application [17]. In hearing aids, microphones are rarely matched in gain and phase. In [3], gain and phase differences between microphone characteristics of up to 6 dB and 10°, respectively, have been reported.

4.3.2 Comparison of different FD stochastic gradient techniques

Figure 9(a) and (b) compare the performance of the different FD Stochastic Gradient (SG) SP-SDW-MWF algorithms without $\mathbf{w}_0$ (i.e., the SDR-GSC) as a function of the trade-off parameter μ for a stationary and non-stationary (e.g., multi-talker babble) noise source, respectively, at 90°. To analyze the impact of the approximation (64) on the performance, the result of a FD implementation of (63), which uses the clean speech, is depicted too. For both noise scenarios, the stochastic gradient algorithm significantly outperforms the NLMS based algorithm, especially for $\frac{1}{\mu} \rightarrow 1$. Without Low Pass (LP) filter, both algorithms achieve a worse improvement compared to (63), especially for large μ. For a stationary speech-like noise source, the FD-SG algorithm does not suffer too much from approximation (64). In a highly time-varying noise

scenario, such as multi-talker babble, the limited averaging of $\mathbf{r}[k]$ in the FD implementation does not suffice to maintain the large noise reduction achieved by (63). The loss in noise reduction performance could be reduced by decreasing the step-size ρ', at the expense of a reduced convergence speed. Applying the low pass filter (95) significantly improves the performance for all $\frac{1}{\mu}$, while changes in the noise scenario can still be tracked.

Figure 10 plots the improvement $\Delta\text{SNR}_{\text{intellig}}$ and $\text{SD}_{\text{intellig}}$ of the SP-SDW-MWF ($\frac{1}{\mu} = 0.5$) with and without filter $\mathbf{w}_0$ for the babble noise scenario as a function of $\frac{1}{1-\lambda}$ where λ is the exponential weighting factor of the LP filter (see (95)). Performance clearly improves for increasing λ. For small λ, the SP-SDW-MWF with $\mathbf{w}_0$ suffers from a larger excess error -and hence worse $\Delta\text{SNR}_{\text{intellig}}$- compared to the SP-SDW-MWF without $\mathbf{w}_0$. This is due to the larger dimensions of $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$.

The LP filter avoids that the desired speech is distorted by a highly time-varying filter $\mathbf{w}_s$. In contrast to a decrease in step size ρ', the LP filter does not compromise tracking of changes in the noise scenario. As an illustration, Figure 11 plots the convergence behavior of the FD stochastic gradient algorithm without $\mathbf{w}_0$ (i.e., the SDR-GSC) for $\lambda = 0$ and $\lambda = 0.9998$, respectively, when the noise source position suddenly changes from 90° to 180°. A gain mismatch Υ_2 of 4 dB was applied to the second microphone. To avoid fast fluctuations in the residual noise energy ϵ_n^2 and speech distortion energy ϵ_d^2, the desired and interfering noise source in this experiment are stationary, speech-like. The upper figure depicts the residual noise energy ϵ_n^2 as a function of the number of input samples, the lower figure plots the residual speech distortion ϵ_d^2 during speech + noise periods as a function of the number of speech + noise samples. Both algorithms (i.e., $\lambda = 0$ and $\lambda = 0.9998$) have about the same convergence rate. When the change in position occurs, the algorithm with $\lambda = 0.9998$ even converges faster. For $\lambda = 0$, the approximation error (64) remains large for a while since the noise vectors in the buffer are not up to date. For $\lambda = 0.9998$, the impact of the instantaneous large approximation error is reduced thanks to the low pass filter.

4.3.3 Comparison with SPA

Figure 12 and Figure 13 compare the performance of the FD stochastic gradient algorithm with LP filter ($\lambda = 0.9998$) and the FD-NLMS based SPA in a multiple noise source scenario. The noise scenario consists of 5 multi-talker babble noise sources positioned at angles 75°, 120°, 180°, 240°, 285° w.r.t. the desired source at 0°. To assess the sensitivity of the algorithms against errors in the assumed signal model, the influence of microphone mismatch, e.g., gain mismatch $\Upsilon_2 = 4$ dB of the second microphone, on the performance is depicted too. In Figure 12, the improvement $\Delta\text{SNR}_{\text{intellig}}$ and the distortion $\text{SD}_{\text{intellig}}$ of the SP-SDW-MWF with and without filter $\mathbf{w}_0$ is depicted as a function of the trade off factor $\frac{1}{\mu}$. Figure 13 shows the results of the QIC-GSC

$$\mathbf{w}^H\mathbf{w} \leq \beta^2 \tag{99}$$

for different constraint values β^2, which is implemented using the FD-NLMS based SPA.

Both, the SPA and the stochastic gradient based SP-SDW-MWF increase the robustness of the GSC (i.e., the SP-SDW-MWF without $\mathbf{w}_0$ and $\frac{1}{\mu} = 0$). For a given maximum allowable distortion $\text{SD}_{\text{intellig}}$,

the SP-SDW-MWF with and without $\mathbf{w}_0$ achieve a better noise reduction performance than the SPA. The performance of the SP-SDW-MWF with $\mathbf{w}_0$ is -in contrast to the SP-SDW-MWF without $\mathbf{w}_0$- not affected by microphone mismatch. In the absence of model errors, the SP-SDW-MWF with $\mathbf{w}_0$ achieves a slightly worse performance than the SP-SDW-MWF without $\mathbf{w}_0$. With $\mathbf{w}_0$, the estimate of $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ is less accurate due to the larger dimensions of $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ (see also Figure 10).

In short, the proposed stochastic gradient implementation of the SP-SDW-MWF preserves the benefit of the SP-SDW-MWF over the QIC-GSC.

4.4 Conclusions

In this paper, we derived time-domain and frequency-domain stochastic gradient algorithms for the SP-SDW-MWF and compared their performance to the SPA. Starting from the cost function of the SP-SDW-MWF, a time-domain stochastic gradient algorithm has been derived in Section 4.1. In addition, the LMS based algorithm [26] has been extended so that it applies to the SP-SDW-MWF. To increase convergence and reduce complexity, a frequency-domain implementation has been proposed. Both, the stochastic gradient and LMS based algorithm suffer from a large excess error when applied in highly time-varying noise scenarios. In Section 4.2, we show that the excess error is reduced by applying a low pass filter to the part of the gradient estimate that limits speech distortion. The low pass filtering avoids a highly time-varying distortion of the desired speech component while not degrading the tracking performance needed in time-varying noise scenarios. Section 4.3 compares the performance of the different frequency-domain stochastic gradient algorithms for a hearing aid application. The stochastic gradient SP-SDW-MWF outperforms the LMS based algorithm, while complexity is not increased. For a non-stationary noise scenario, the LMS based and stochastic gradient SP-SDW-MWF suffer from a reasonably large excess error. Experimental results show that the low pass filtering significantly improves the performance of the stochastic gradient algorithm and does not compromise the tracking of changes in the noise scenario. In addition, experiments demonstrate that the proposed stochastic gradient algorithm preserves the benefit of the SP-SDW-MWF over QIC-GSC. The limited computational cost and the better noise reduction performance of the proposed algorithm make it a good alternative to the SPA for implementation in hearing aids.

References

[1] R. W. Stadler and W. M. Rabinowitz, "On the potential of fixed arrays for hearing aids," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1332–1342, Sept. 1993.

[2] P. M. Peterson, *Adaptive array processing for multiple microphone hearing aids*, Ph.D. thesis, Dept. Elect. Eng. and Comp. Sci., M.I.T, Cambridge, MA, 1989, available as Res. Lab. Elect. Techn. Rept. 541.

[3] L. B. Jensen, "Hearing Aid with adaptive Matching of Input Transducers," U.S. patent 0041696, Apr. 2002.

[4] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[5] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[6] K. M. Buckley, "Broad-Band Beamforming and the Generalized Sidelobe Canceller," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, no. 5, pp. 1322–1323, Oct. 1986.

[7] J. E. Greenberg and P. M. Zurek, "Evaluation of an Adaptive Beamforming Method for Hearing Aids," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1662–1676, Mar. 1992.

[8] J. Vanden Berghe and J. Wouters, "An adaptive noise canceller for hearing aids using two nearby microphones," *J. Acoust. Soc. Amer.*, vol. 103, pp. 3621–3626, June 1998.

[9] M. W. Hoffman and K. M. Buckley, "Robust time-domain processing of broadband acoustic data," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 193–203, May 1995.

[10] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix Using Constrained Adaptive Filters," *IEEE Trans. Signal Processing*, vol. 47, pp. 2677–2683, 1999.

[11] F. Luo, J. Yang, C. Pavlovic, and A. Nehorai, "Adaptive Null-Forming Scheme in Digital Hearing Aids," *Signal Processing*, vol. 50, no. 7, pp. 1583–1590, July 2002.

[12] B. Widrow and S. Stearns, *Adaptive Signal Processing*, Prentice Hall, Englewood Cliffs, 1985.

[13] D. Van Compernolle, "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, Apr. 1990, vol. 2, pp. 833–836.

[14] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust Adaptive Beamforming," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[15] N. K. Jablon, "Adaptive beamforming with the generalized sidelobe can in the presence of array imperfections," *IEEE Trans. Antennas Propag.*, vol. 34, pp. 996–1012, Aug. 1986.

[16] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of GSVD based optimal Filtering and generalized Sidelobe Canceller for Hearing Aid Applications," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2001, pp. 31–34.

[17] A. Spriet, M. Moonen, and J. Wouters, "Robustness Analysis of Multi-channel Wiener Filtering and Generalized Sidelobe Cancellation for Multi-microphone Noise Reduction in Hearing Aid Applications," Tech. Rep. ESAT-SISTA/TR 02-81, ESAT-SCD/SISTA, KU Leuven (Belgium), Sept. 2002, available at ftp://ftp.esat.kuleuven.ac.be/sista/spriet/reports/02-81.pdf.

[18] M. W. Hoffman, T. D. Trine, K. M. Buckley, and D. J. Van Tasell, "Robust Adaptive microphone array processing for hearing aids: realistic speech enhancement predictions," *J. Acoust. Soc. Amer.*, vol. 96, pp. 759–770, 1994.

[19] Z. Tian, K.L. Bell, and H.L. Van Trees, "A Recursive Least Squares Implementation for LCMP Beamforming Under Quadratic Constraint," *IEEE Trans. Signal Processing*, vol. 49, no. 6, pp. 1138–1145, June 2001.

[20] S. Doclo and M. Moonen, "GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[21] S. Doclo and M. Moonen, *GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement*, chapter 6 in "Microphone Arrays: Signal Processing Techniques and Applications" (Brandstein, M. S. and Ward, D. B., Eds.), pp. 111–132, Springer-Verlag, May 2001.

[22] G. Rombouts and M. Moonen, "QRD-based optimal filtering for acoustic noise reduction," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Toulouse, France, Sept. 2002, vol. 3, pp. 301–304.

[23] A. Spriet, M. Moonen, and J. Wouters, "A multi-channel subband generalized singular value decomposition approach to speech enhancement," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 149–158, Mar.-Apr. 2002.

[24] A. Spriet, M. Moonen, and J. Wouters, "A multichannel subband GSVD based approach for speech enhancement in hearing aids," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sept. 2001, pp. 187–191.

[25] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Trans. Speech, Audio Processing*, vol. 7, no. 3, pp. 241–22, May 1999.

[26] D. A. Florêncio and H. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, May 2001.

[27] S. Doclo and M. Moonen, "Noise Reduction in Multi-Microphone Speech Signals using Recursive and Approximate GSVD-based Optimal Filtering," in *Proc. of the IEEE Benelux Signal Processing Symposium (SPS2000)*, Hilvarenbeek, The Netherlands, Mar. 2000.

[28] A. Spriet, M. Moonen, and J. Wouters, "A multi-channel subband gsvd approach to speech enhancement," *ETT*, vol. 13, no. 2, pp. 149–158, 2002.

[29] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," Tech. Rep. ESAT-SISTA/TR 03-46, ESAT/SISTA, K.U. Leuven (Belgium), 2003.

[30] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[31] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," Accepted for publication in IEEE Transactions on Signal Processing, Jan. 2003. Available at ftp://ftp.esat.kuleuven.ac.be/sista/doclo/reports/02-111.ps.gz.

[32] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 251–266, July 1995.

[33] N. Grbic and S. Nordholm, "Soft contrained subband beamforming for hands-free speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002, pp. 885–888.

[34] S. Nordebo, I. Claesson, and S. Nordholm, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE Journal of Oceanic Engineering*, vol. 19, no. 4, pp. 583–590, Oct. 1994.

[35] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, vol. 34, pp. 3–12, Apr. 2001.

[36] International Collegium of Rehabilitative Audiology, "Noise Signals ICRA Compact Disc," Ver 0.3 1997.

[37] M. J. Link and K. M. Buckley, "Prewhitening for intelligibility gain in hearing aid arrays," *J. Acoust. Soc. Amer.*, vol. 93, no. 4, pp. 2139–2140, Apr. 1993.

[38] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[39] Acoustical Society of America, "ANSI S3.5-1997 American National Standard Methods for Calculation of the Speech Intelligibility Index," June 1997.

[40] B. Widrow, J.M. McCool, M.G. Larimore, and C.R. Johnson Jr., "Stationary and nonstationary learning characteristics of the lms adaptive filter," *Proceedings of IEEE*, vol. 16, pp. pp. 1151–1162, 1976.

Multi-microphone adaptive noise reduction techniques for speech enhancement

Addendum 1

Simon Doclo[1], Ann Spriet[1,2], Marc Moonen[1], Jan Wouters[2]

March 15, 2004

[1]Katholieke Universiteit Leuven - ESAT/SCD
Kasteelpark Arenberg 10
B-3001 Leuven-Heverlee, Belgium
Tel. +32 16 32 17 95 Fax +32 16 32 19 70
{doclo,spriet,moonen}@esat.kuleuven.ac.be

[2]Katholieke Universiteit Leuven
ENT-Dept./Lab Exp ORL
Kapucijnenvoer 33
B-3000 Leuven, Belgium
jan.wouters@uz.kuleuven.ac.be

A Efficient frequency-domain implementation using correlation matrices

In Section 4 stochastic gradient algorithms in the time-domain and in the frequency-domain have been developed for implementing the Speech Distortion Weighted Multi-channel Wiener Filter (SDW-MWF). These algorithms however require large data buffers for calculating the regularisation term required in the filter update formulas, resulting in a large memory usage. In this addendum it is shown that by approximating this regularisation term in the frequency-domain, (diagonal) speech and noise correlation matrices need to be stored instead of data buffers, such that the memory usage is decreased drastically, while also the computational complexity is further reduced. Experimental results demonstrate that this approximation results in a small - positive or negative - performance difference, such that the proposed algorithm preserves the robustness benefit of the SP-SDW-MWF over the QIC-GSC, while both its computational complexity and memory usage are now comparable to the NLMS-based SPA for implementing the QIC-GSC.

A.1 Stochastic Gradient algorithms

In this section we first briefly review the stochastic gradient algorithm in the time-domain (cf. Section 4.1.1) and the calculation of the regularisation term $\mathbf{r}[k]$ (cf. Sections 4.1.1 and 4.2.1). We then show that by approximating this regularisation term in the frequency-domain the memory usage can be reduced drastically..

A.1.1 Time-Domain implementation

In Section 4.1.1 a stochastic gradient algorithm in the time-domain has been developed for minimising the cost function in (42), i.e.

$$\mathbf{w}[k+1] = \mathbf{w}[k] + \rho \left\{ \mathbf{y}^n[k](y_0^{n,*}[k-\Delta] - \mathbf{y}^{n,H}[k]\mathbf{w}[k]) - \mathbf{r}[k] \right\} \quad (109)$$

$$\mathbf{r}[k] = \frac{1}{\mu}\mathbf{y}^s[k]\mathbf{y}^{s,H}[k]\mathbf{w}[k] \quad (110)$$

$$\rho = \frac{\rho'}{\mathbf{y}^{n,H}[k]\mathbf{y}^n[k] + \frac{1}{\mu}\mathbf{y}^{s,H}[k]\mathbf{y}^s[k] + \delta}, \quad (111)$$

with ρ the normalised step size of the adaptive algorithm, δ a small positive constant, and $\mathbf{w}[k]$, $\mathbf{y}^n[k]$, $\mathbf{y}^s[k]$ and $\mathbf{r}[k]$ NL-dimensional vectors. For $1/\mu = 0$ and no filter $\mathbf{w}_0$ present, (109) reduces to an NLMS-type update formula often used in GSC, and *operated during noise-only-periods* [7, 10, 13, 42]. For $1/\mu \neq 0$, the additional regularisation term $\mathbf{r}[k]$ limits speech distortion due to possible signal model errors.

In order to compute (110), knowledge about the (instantaneous) correlation matrix $\mathbf{y}^s[k]\mathbf{y}^{s,H}[k]$ of the clean speech signal is required, which is obviously not available. In order to avoid the need for calibration, it is suggested in Section 4.1.1 to store L-dimensional speech+noise-vectors $\mathbf{y}_i[k]$, $i = M-N \ldots M-1$, during speech-periods in a circular *speech+noise-buffer* $\mathbf{B}_1 \in \mathbb{R}^{N \times L_{buf1}}$ and to adapt the filter $\mathbf{w}[k]$ using

(109) during *noise-only-periods*[20], based on approximating the regularisation term in (110) by

$$\mathbf{r}[k] = \frac{1}{\mu}\left(\mathbf{y}_{buf_1}[k]\mathbf{y}_{buf_1}^H[k] - \mathbf{y}^n[k]\mathbf{y}^{n,H}[k]\right)\mathbf{w}[k]\,, \tag{112}$$

with $\mathbf{y}_{buf_1}[k]$ a vector from the circular speech+noise-buffer $\mathbf{B_1}$, cf. (72). However, as has been indicated in Section 4.1.1, this estimate of $\mathbf{r}[k]$ is quite poor, resulting in a large excess error, especially for small μ and large ρ'. Hence, it has been suggested to use an estimate of the average clean speech correlation matrix $\mathcal{E}\{\mathbf{y}^s[k]\mathbf{y}^{s,H}[k]\}$ in (110), such that $\mathbf{r}[k]$ can be computed as

$$\mathbf{r}[k] = \frac{1}{\mu}(1-\tilde{\lambda})\sum_{l=0}^{k}\tilde{\lambda}^{k-l}\left(\mathbf{y}_{buf_1}[l]\mathbf{y}_{buf_1}^H[l] - \mathbf{y}^n[l]\mathbf{y}^{n,H}[l]\right)\cdot\mathbf{w}[k]\,, \tag{113}$$

with $\tilde{\lambda}$ an exponential weighting factor and the step size ρ in (111) now equal to

$$\rho = \frac{\rho'}{\mathbf{y}^{n,H}[k]\mathbf{y}^n[k] + \frac{1}{\mu}(1-\tilde{\lambda})\sum_{l=0}^{k}\tilde{\lambda}^{k-l}\left|\mathbf{y}_{buf_1}^H[l]\mathbf{y}_{buf_1}[l] - \mathbf{y}^{n,H}[l]\mathbf{y}^n[l]\right| + \delta}\,.$$

For *stationary noise* a small $\tilde{\lambda}$, i.e. $1/(1-\tilde{\lambda}) \sim NL$, suffices. However, in practice the speech and the noise signals are often *spectrally highly non-stationary* (e.g. multi-talker babble noise), whereas their *long-term* spectral and spatial characteristics usually vary more slowly in time. Spectrally highly non-stationary noise can still be spatially suppressed by using an estimate of the *long-term* correlation matrix in $\mathbf{r}[k]$, i.e. $1/(1-\tilde{\lambda}) \gg NL$.

In order to avoid expensive matrix operations for computing (113), it is assumed in Section 4.2.1 that $\mathbf{w}[k]$ varies slowly in time, i.e. $\mathbf{w}[k] \approx \mathbf{w}[l]$, such that (113) can be approximated with vector instead of matrix operations by directly applying a low-pass filter to the regularisation term $\mathbf{r}[k]$, cf. (100),

$$\mathbf{r}[k] = \frac{1}{\mu}(1-\tilde{\lambda})\sum_{l=0}^{k}\tilde{\lambda}^{k-l}\left(\mathbf{y}_{buf_1}[l]\mathbf{y}_{buf_1}^H[l] - \mathbf{y}^n[l]\mathbf{y}^{n,H}[l]\right)\cdot\mathbf{w}[l] \tag{114}$$

$$= \tilde{\lambda}\mathbf{r}[k-1] + (1-\tilde{\lambda})\frac{1}{\mu}\left(\mathbf{y}_{buf_1}[k]\mathbf{y}_{buf_1}^H[k] - \mathbf{y}^n[k]\mathbf{y}^{n,H}[k]\right)\mathbf{w}[k]\,. \tag{115}$$

However, as will be shown in the next paragraph, this assumption is actually not required in a frequency-domain implementation.

A.1.2 Efficient Frequency-Domain implementation

In Section 4.2.2 the (improved) stochastic gradient algorithm in the time-domain has been converted to a frequency-domain implementation by using a block-formulation and overlap-save procedures (similar

[20] In Section 4.1.1 it has been shown that storing noise-only-vectors $\mathbf{y}_i[k] = \mathbf{y}_i^n[k]$, $i = 0 \ldots M-1$ during noise-only-periods in a circular *noise-buffer* $\mathbf{B_2} \in \mathbb{R}^{M \times L_{buf2}}$ additionally allows adaptation during speech-periods.

to standard adaptive filtering techniques in the frequency-domain [43]). However, the frequency-domain algorithm described in Section 4.2.2 (**Algorithm 3**) requires many data buffers and hence the storage of a large amount of data[21]. A substantial memory (and computational complexity) reduction can be achieved by the following two steps:

- When using (113) instead of (115) for calculating the regularisation term, *correlation matrices* instead of data samples need to be stored. The frequency-domain implementation of the resulting algorithm is then summarised in **Algorithm 4**, where $2L \times 2L$-dimensional speech and noise correlation matrices $\mathbf{S}_{ij}[k]$ and $\mathbf{S}^n_{ij}[k]$, $i,j = M-N \ldots M-1$ are used for calculating the regularisation term $\mathbf{R}_i[k]$ and (part of) the step size $\mathbf{\Lambda}[k]$. These correlation matrices are updated respectively during speech-periods and noise-only-periods[22]. However, this first step does not necessarily reduce the memory usage (NL_{buf_1} for data buffers vs. $2(NL)^2$ for correlation matrices) and will even increase the computational complexity, since the correlation matrices are not diagonal.

- The correlation matrices in the frequency-domain can be approximated by diagonal matrices, since $\mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}$ in Algorithm 4 can be well approximated by $\mathbf{I}_{2L}/2$ [44, 45]. Hence, the speech and the noise correlation matrices are updated as

$$\mathbf{S}_{ij}[k] = \lambda \mathbf{S}_{ij}[k-1] + (1-\lambda)\mathbf{Y}_i^H[k]\mathbf{Y}_j[k]/2\,, \tag{116}$$

$$\mathbf{S}^n_{ij}[k] = \lambda \mathbf{S}^n_{ij}[k-1] + (1-\lambda)\mathbf{Y}_i^{n,H}[k]\mathbf{Y}^n_j[k]/2\,, \tag{117}$$

  leading to a significant reduction in memory usage and computational complexity, cf. Section A.2, while having a minimal impact on the performance and the robustness, cf. Section A.3. We will refer to this algorithm as **Algorithm 5**. Algorithm 5 is in fact quite similar to the algorithm presented in [46], which is derived directly from a frequency-domain cost function. Some major differences however exist, e.g. in [46] the regularisation term $\mathbf{R}_i[k]$ is absent, the term $\mathbf{F}\mathbf{g}\mathbf{F}^{-1}$ is also approximated by $\mathbf{I}_{2L}/2$ and the speech and the noise correlation matrices are block-diagonal.

A.2 Memory usage and computational complexity

Table 2 summarises the computational complexity and the memory usage of the frequency-domain NLMS-based SPA for implementing the QIC-GSC [14][23] and the frequency-domain stochastic gradient algorithms for implementing the SDW-MWF (Algorithm 3 and Algorithm 5). As in Section 4.2.3, the computational complexity is expressed as the number of operations per second (MIPS), while the memory usage is expressed in kWords. We assume that one complex multiplication is equivalent to 4 real multiplications and 2

[21] In order to achieve a good performance, typical values for the buffer lengths L_{buf_1} and L_{buf_2} of the circular buffers $\mathbf{B}_1$ and $\mathbf{B}_2$ are 10000...20000.

[22] When using correlation matrices, filter adaptation can only take place during noise-only-periods, since during speech-periods the desired signal $\mathbf{d}[k]$ cannot be constructed from the noise-buffer $\mathbf{B}_2$ any more.

[23] The computational complexity of the frequency-domain QIC-GSC using SPA also represents the complexity when the adaptive filter is only updated during noise-only-periods.

Algorithm 4 Frequency-domain implementation with correlation matrices (without approximation)

Initialisation and matrix definitions:

$\mathbf{W}_i[0] = [\ 0 \ \cdots \ 0\ ]^T, i = M-N \ldots M-1$

$P_m[0] = \delta_m, m = 0 \ldots 2L-1$

$\mathbf{F} = 2L \times 2L$-dimensional DFT matrix

$\mathbf{g} = \begin{bmatrix} \mathbf{I}_L & \mathbf{0}_L \\ \mathbf{0}_L & \mathbf{0}_L \end{bmatrix}, \quad \mathbf{k} = [\ \mathbf{0}_L \ \ \mathbf{I}_L\ ]$

$\mathbf{0}_L = L \times L$-dimensional matrix with zeros, $\mathbf{I}_L = L \times L$-dimensional identity matrix

For each new block of L samples (per channel):

$\mathbf{d}[k] = [\ y_0[kL-\Delta] \ \cdots \ y_0[kL-\Delta+L-1]\ ]^T$

$\mathbf{Y}_i[k] = \mathrm{diag}\left\{\mathbf{F}\,[\ y_i[kL-L] \ \cdots \ y_i[kL+L-1]\ ]^T\right\}, i = M-N \ldots M-1$

Output signal:

$$\mathbf{e}[k] = \mathbf{d}[k] - \mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j[k]\mathbf{W}_j[k], \quad \mathbf{E}[k] = \mathbf{F}\mathbf{k}^T\mathbf{e}[k]$$

If speech detected:

$$\mathbf{S}_{ij}[k] = (1-\lambda)\sum_{l=0}^{k}\lambda^{k-l}\mathbf{Y}_i^H[l]\mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\mathbf{Y}_j[l] = \lambda\mathbf{S}_{ij}[k-1] + (1-\lambda)\mathbf{Y}_i^H[k]\mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\mathbf{Y}_j[k]$$

If noise detected: $\mathbf{Y}_i[k] = \mathbf{Y}_i^n[k]$

$$\mathbf{S}_{ij}^n[k] = (1-\lambda)\sum_{l=0}^{k}\lambda^{k-l}\mathbf{Y}_i^{n,H}[l]\mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\mathbf{Y}_j^n[l] = \lambda\mathbf{S}_{ij}^n[k-1] + (1-\lambda)\mathbf{Y}_i^{n,H}[k]\mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\mathbf{Y}_j^n[k]$$

Update formula (only during noise-only-periods):

$$\mathbf{R}_i[k] = \frac{1}{\mu}\sum_{j=M-N}^{M-1}\left[\mathbf{S}_{ij}[k] - \mathbf{S}_{ij}^n[k]\right]\mathbf{W}_j[k], i = M-N \ldots M-1$$

$$\mathbf{W}_i[k+1] = \mathbf{W}_i[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\mathbf{\Lambda}[k]\left\{\mathbf{Y}_i^{n,H}[k]\mathbf{E}[k] - \mathbf{R}_i[k]\right\}, i = M-N \ldots M-1$$

with

$$\mathbf{\Lambda}[k] = \frac{2\rho'}{L}\,\mathrm{diag}\,\{P_0^{-1}[k], \ldots, P_{2L-1}^{-1}[k]\}$$

$$P_m[k] = \gamma P_m[k-1] + (1-\gamma)\,(P_{1,m}[k] + P_{2,m}[k]), m = 0 \ldots 2L-1$$

$$P_{1,m}[k] = \sum_{j=M-N}^{M-1}\left|Y_{j,m}^n[k]\right|^2, \ P_{2,m}[k] = \frac{1}{\mu}\left|\sum_{j=M-N}^{M-1}S_{jj,m}[k] - S_{jj,m}^n[k]\right|, m = 0 \ldots 2L-1$$

Algorithm	Computational complexity		MIPS
	update formula	adaptation of step size	
NLMS based SPA (QIC-GSC)	$(14M-11-\frac{4(M-1)}{L})+(6M-2)\log_2 2L$MAC $+1/L$Sq$+1/L$D	$(2M+2)$MAC $+1$D	2.16
SG with LP (Algorithm 3)	$(26N+4-\frac{10N}{L})+(6N+10)\log_2 2L$MAC	$(4N+6)$MAC $+1$D$+1$Abs	3.22[a], 4.27[b]
SG with LP (Algorithm 5)	$(10N^2+13N-\frac{4N^2+3N}{L})+(6N+4)\log_2 2L$MAC	$(2N+4)$MAC $+1$D$+1$Abs	2.71[a], 4.31[b]
	Memory usage		kWords
QIC-GSC	$4(M-1)L+6L$		0.45
Algorithm 3	$2NL_{buf_1}+6LN+7L$		40.61[a], 60.80[b]
Algorithm 5	$4LN^2+6LN+7L$		1.12[a], 1.95[b]

Table 2: Computational complexity and memory usage for $M=3$, $L=32$, $f_s=16$ kHz, $L_{buf_1}=10000$, (a) $N=M-1$, (b) $N=M$

real additions and that a $2L$-point FFT of a real input vector requires $2L\log_2 2L$ real MACs (assuming the radix-2 FFT algorithm). From this table we can draw the following conclusions:

- The *computational complexity* of the SDW-MWF (Algorithm 3) with filter $\mathbf{w}_0$ is about twice the complexity of the QIC-GSC (and even less if the filter $\mathbf{w}_0$ is not present). The approximation of the regularisation term in Algorithm 5 further reduces the computational complexity. However, this only remains true for a small number of input channels, since the approximation introduces a quadratic term $\mathcal{O}(N^2)$.

- Due to the storage of the data samples used in the circular speech+noise-buffer $\mathbf{B}_1$, the *memory usage* of the SDW-MWF (Algorithm 3) is quite high in comparison with the QIC-GSC (depending on the size of the data buffer L_{buf_1} of course). By using the approximation of the regularisation term in Algorithm 5, the memory usage can be reduced drastically, since now diagonal correlation matrices instead of data buffers need to be stored. Note however that also for the memory usage a quadratic term $\mathcal{O}(N^2)$ is present.

A.3 Experimental results

In this paragraph it is shown that practically no performance difference exists between Algorithm 3 and Algorithm 5, such that the SDW-MWF using the implementation proposed in this addendum indeed preserves its robustness benefit over the GSC (and the QIC-GSC).

A.3.1 Set-up

The same set-up has been used as in Section 4.3.1. A 3-microphone BTE hearing aid with 3 omnidirectional microphones (Knowles FG-3452) has been mounted on a dummy head in an office room. The interspacing d between the first and the second microphone is about $d = 1$ cm and the interspacing between the second and the third microphone is about 1.5 cm. The reverberation time $T_{60\,\mathrm{dB}}$ is about 700 ms for a speech weighted noise. The desired speech source and the noise sources are positioned at a distance of 1 m from the head. The desired speech source is positioned in front of the head (at 0°) and consists of English sentences. The noise scenario consists of five multi-talker babble noise sources, positioned at 75°, 120°, 180°, 240° and 285°. The desired signal and the total noise signal both have a level of 70 dB SPL at the centre of the head. For evaluation purposes, the speech and the noise signals have been recorded separately.

The microphone signals are pre-whitened prior to processing to improve intelligibility [38], and the output is accordingly de-whitened. In the experiments, the microphones have been calibrated by means of recordings of an anechoic speech weighted noise signal positioned at 0° measured while the BTE was mounted on the head. A delay-and-sum beamformer is used as the fixed beamformer and the blocking matrix pairwise subtracts the time-aligned calibrated microphone signals.

The performance of the stochastic gradient algorithms in the frequency-domain is evaluated for a filter length $L = 32$ per channel, $\rho' = 0.8$, $\gamma = 0.95$ and $\lambda = 0.9998$. For all considered algorithms, filter adaptation only takes place during *noise-only periods*. To exclude the effect of the spatial pre-processor, the performance measures are calculated with respect to the output of the fixed beamformer. The sensitivity of the algorithms against errors in the assumed signal model is illustrated for microphone mismatch, i.e. a gain mismatch $\Upsilon_2 = 4$ dB at the second microphone.

A.3.2 Experimental results

Figures 14 and 15 depict the SNR improvement $\Delta\mathrm{SNR}_{\mathrm{intellig}}$ and the speech distortion $\mathrm{SD}_{\mathrm{intellig}}$ of the SP-SDW-MWF (with $\mathbf{w}_0$) and the SDR-GSC (without $\mathbf{w}_0$), implemented using Algorithm 3 (solid line) and Algorithm 5 (dashed line), as a function of the trade-off parameter $1/\mu$. These figures also depict the effect of a gain mismatch $\Upsilon_2 = 4$ dB at the second microphone. From these figures it can be observed that approximating the regularisation term only results in a small performance difference. For most scenarios the performance is even better (i.e. larger SNR improvement and smaller speech distortion) for Algorithm 5 than for Algorithm 3, probably since in Algorithm 3 the additional assumption is used that the filter $\mathbf{w}[k]$ varies slowly in time, cf. (115).

Hence, also when implementing the SDW-MWF using the proposed Algorithm 5, it still preserves its robustness benefit over the GSC (and the QIC-GSC). E.g. it can be observed that the GSC (i.e. SDR-GSC with $1/\mu = 0$) will result in a large speech distortion (and a smaller SNR improvement) when microphone mismatch occurs. Both the SDR-GSC and the SP-SDW-MWF add robustness to the GSC, i.e. the distortion decreases for increasing $1/\mu$. The performance of the SP-SDW-MWF is again hardly affected by microphone mismatch.

A.4 Conclusion

In this addendum we have shown that the memory usage (and the computational complexity) of the SDW-MWF can be reduced drastically by approximating the regularisation term in the frequency-domain, i.e. by computing the regularisation term using (diagonal) frequency-domain correlation matrices instead of time-domain data buffers. It has been shown that approximating the regularisation term only results in a small performance difference, such that the robustness benefit of the SDW-MWF is preserved, while now both the computational complexity and the memory usage are comparable to the NLMS-based SPA for implementing the QIC-GSC.

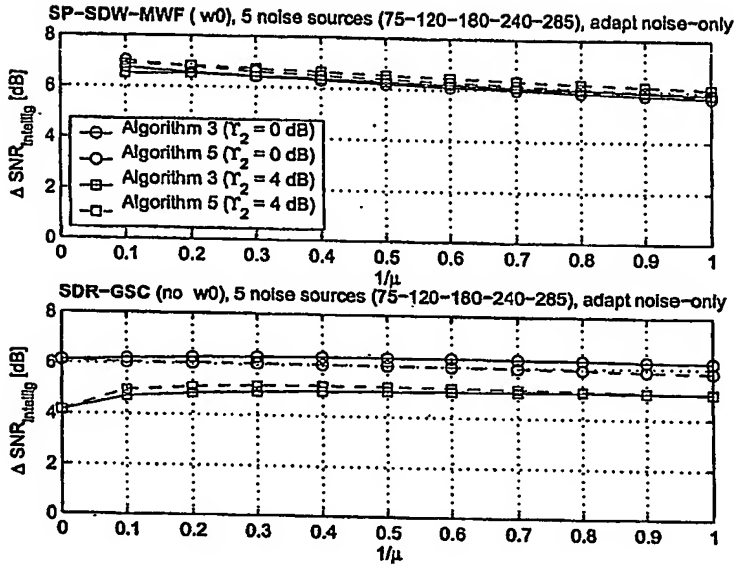

Figure 14: SNR improvement of frequency-domain SP-SDW-MWF (Algorithm 3 and Algorithm 5) in a multiple noise source scenario

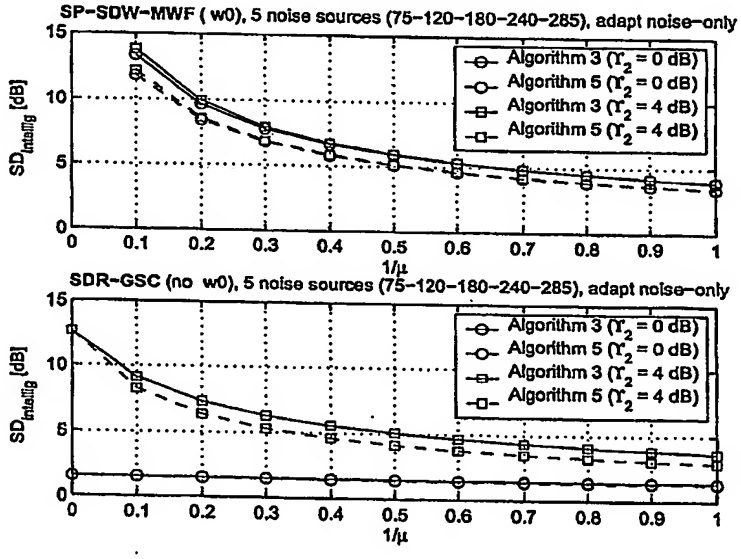

Figure 15: Speech distortion of frequency-domain SP-SDW-MWF (Algorithm 3 and Algorithm 5) in a multiple noise source scenario

[42] S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive Array Noise Suppression of Handsfree Speaker Input in Cars," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 514–518, Nov. 1993.

[43] J. J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering," *IEEE Signal Proc. Magazine*, pp. 15–37, Jan. 1992.

[44] J. Benesty and D. R. Morgan, "Frequency-domain adaptive filtering revisited, generalization to the multi-channel case, and application to acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, May 2000, pp. 789–792.

[45] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *General Derivation of Frequency-Domain Adaptive Filtering*, chapter 8 in "Advances in Network and Acoustic Echo Cancellation", pp. 157–176, Springer-Verlag, 2001.

[46] R. Aichner, W. Herbordt, H. Buchner, and W. Kellermann, "Least-squares error beamforming using minimum statistics and multichannel frequency-domain adaptive filtering," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, Sept. 2003, pp. 223–226.

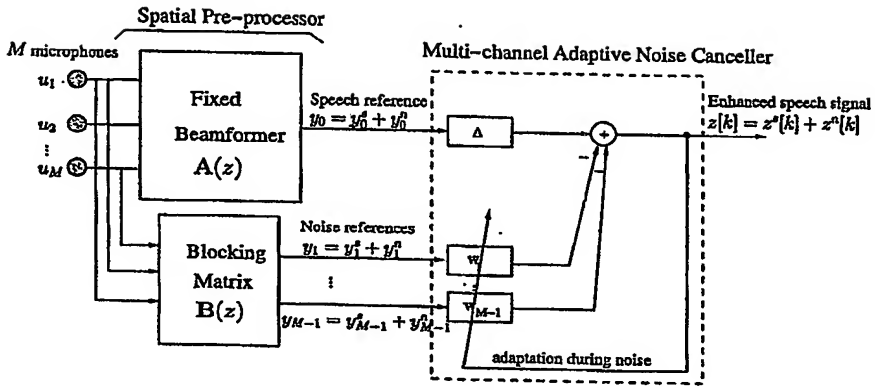

Figure 1: Concept of the Generalized Sidelobe Canceller.

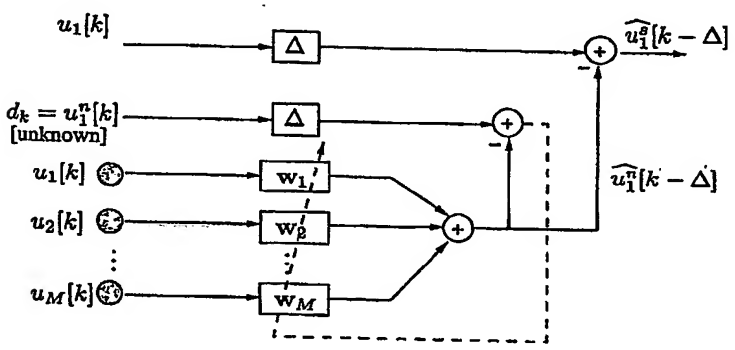

Figure 2: Equivalent approach of multi-channel Wiener filtering.

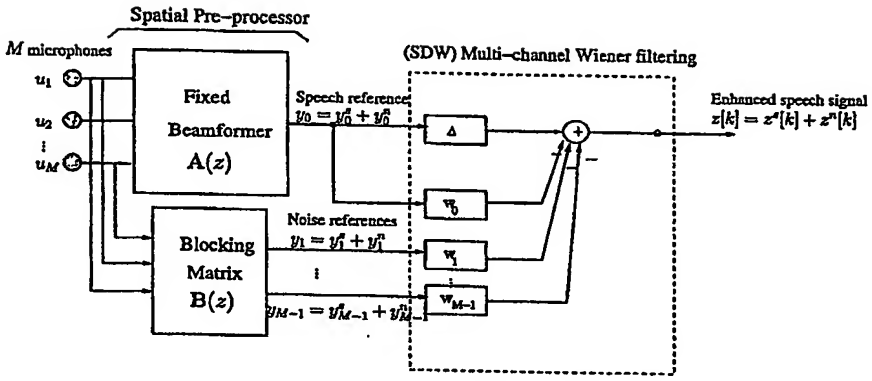

Figure 3: Spatially Pre-processed SDW MWF.

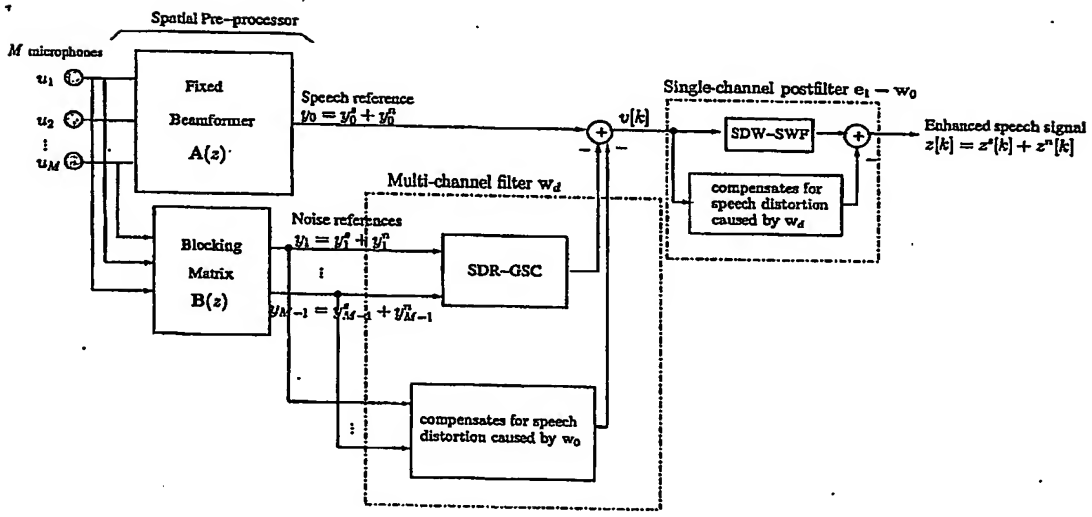

Figure 4: Decomposition of SP-SDW-MWF with $\mathbf{w}_0$ in a multi-channel filter $\mathbf{w}_d$ and single-channel post-filter $\mathbf{e}_1 - \mathbf{w}_0$.

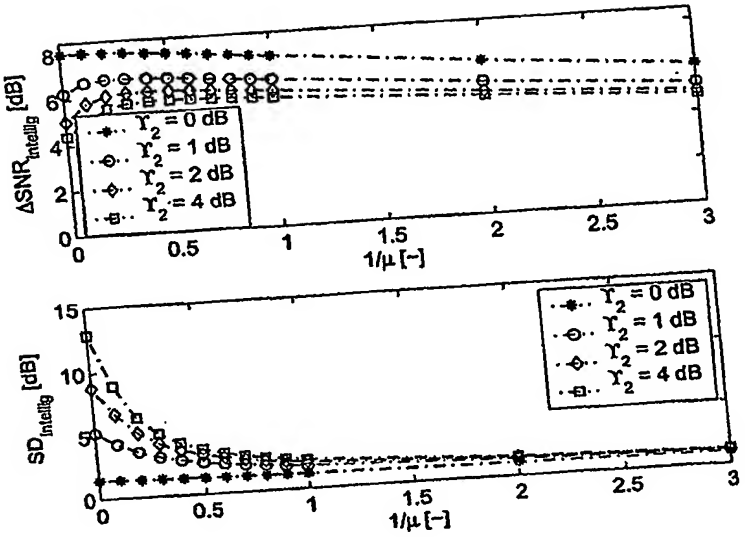

Figure 5: Influence of $1/\mu$ on the performance of the SDR GSC for different gain mismatches Υ_2 at the second microphone.

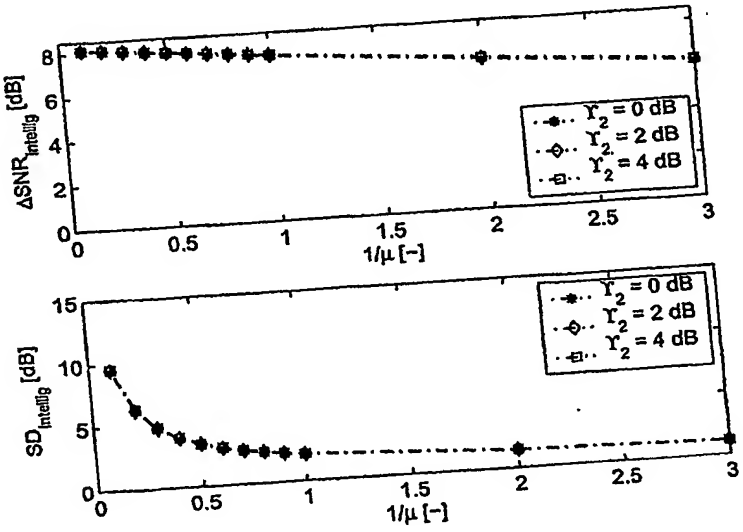

Figure 6: Influence of $1/\mu$ on the performance of the SP SDW MWF with w_0 for different gain mismatches Υ_2 at the second microphone.

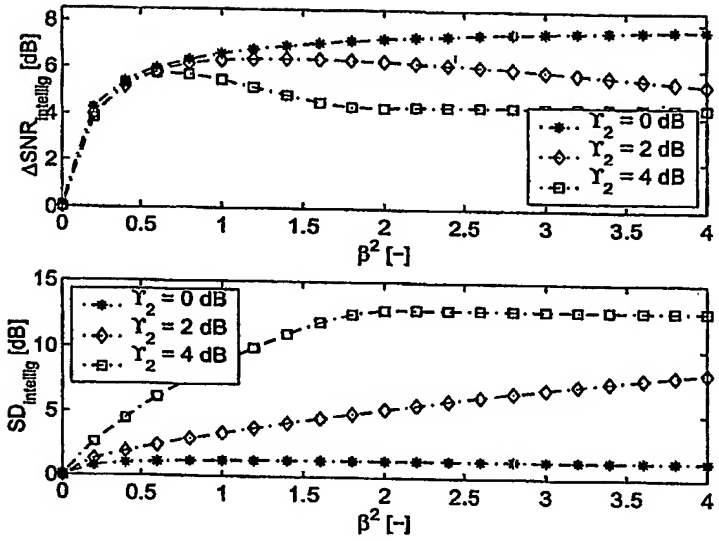

Figure 7: $\Delta SNR_{intellig}$ and $SD_{intellig}$ for QIC-GSC as a function of β^2 for different gain mismatches Υ_2 at the second microphone.

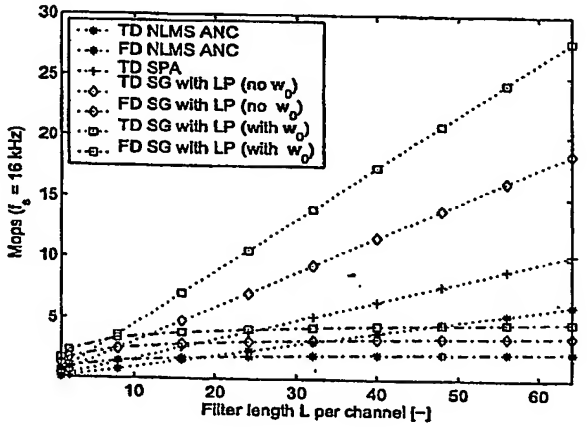

Figure 8: Complexity (expressed in Mops) of TD and FD Stochastic Gradient (SG) algorithm with LP filtering as a function of filter length L per channel; $M = 3$. For comparison, the complexity of the standard NLMS ANC and SPA are depicted too.

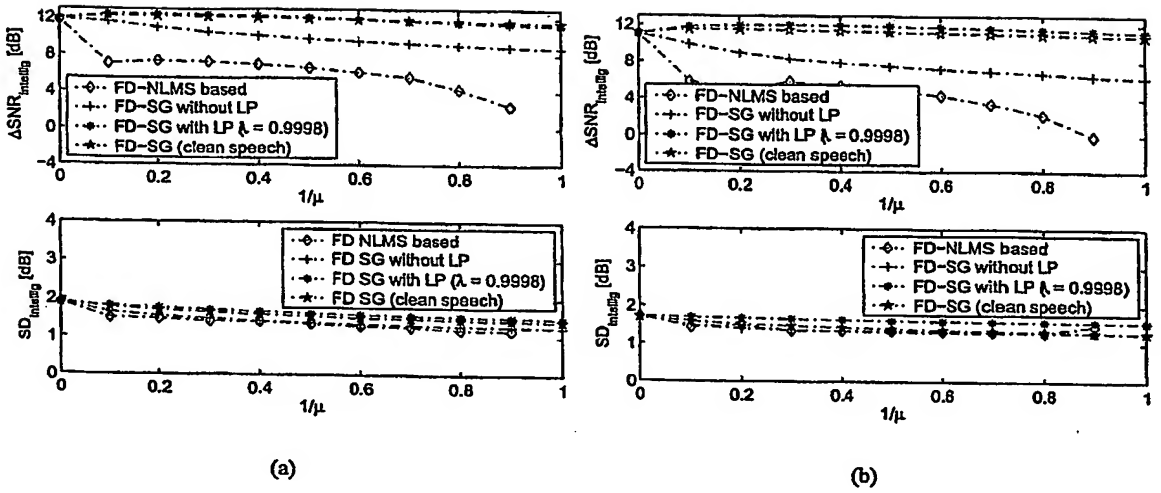

Figure 9: Performance of different FD Stochastic Gradient (FD-SG) algorithms; (a) Stationary speech-like noise at 90°.; (b) Multi-talker babble noise at 90°.

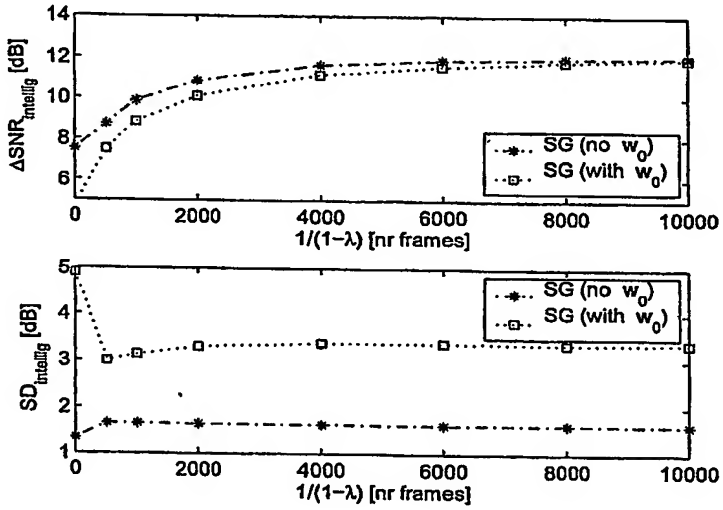

Figure 10: Influence of LP filter on performance of FD stochastic gradient SP-SDW-MWF ($\frac{1}{\mu} = 0.5$) without w_0 and with w_0. Babble noise at 90°.

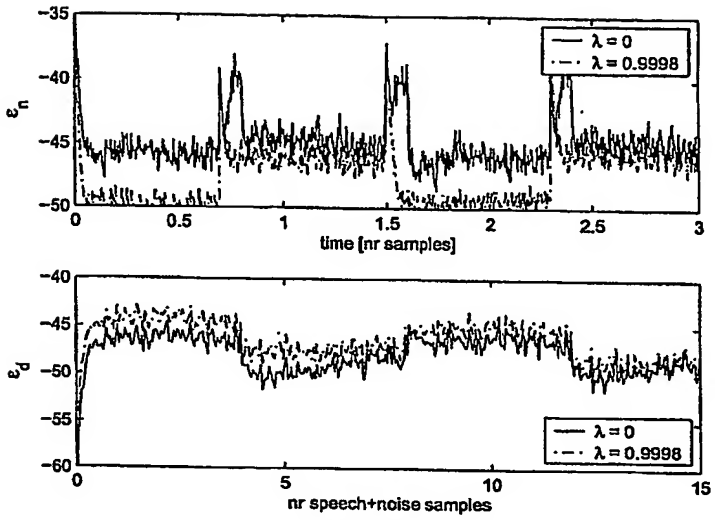

Figure 11: Convergence behavior of FD-SG for $\lambda = 0$ and $\lambda = 0.9998$. The noise source position suddenly changes from 90° to 180° and vice versa.

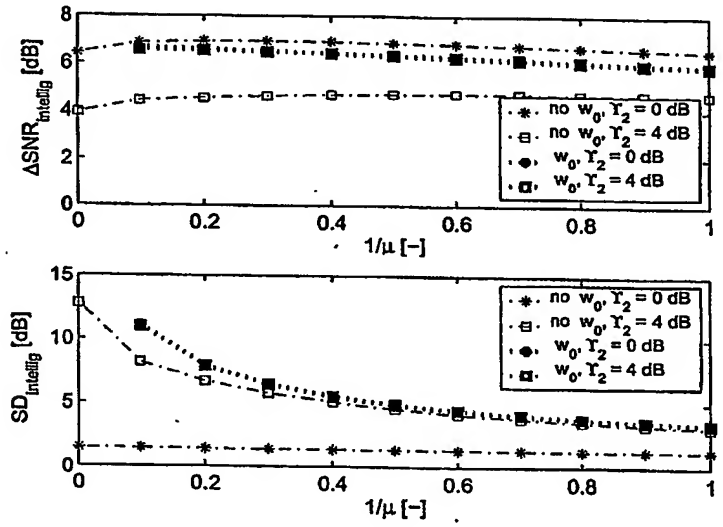

Figure 12: Performance of FD stochastic gradient implementation of SP-SDW-MWF with LP ($\lambda = 0.9998$) in a multiple noise source scenario.

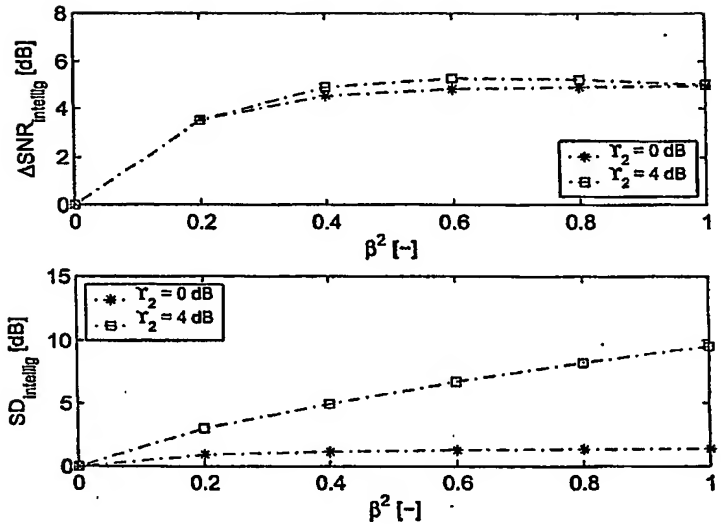

Figure 13: Performance of FD SPA in a multiple noise source scenario.